UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ELVIS NOBREGA DE ALCANTARA

Classificação e aprendizado com proteínas de melanoma

RIO DE JANEIRO
2021

ELVIS NOBREGA DE ALCANTARA

Classificação e aprendizado com proteínas de melanoma

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Daniel Sadoc Menasche
Co-orientador: Prof. Gilberto Barbosa Domont

RIO DE JANEIRO
2021

CIP - Catalogação na Publicação

A347c

Alcantara, Elvis Nobrega de
Classificação e aprendizado com proteínas de melanoma / Elvis Nobrega de Alcantara. -- Rio de Janeiro, 2021.
66 f.

Orientador: Daniel Sadoc Menasche.
Coorientador: Gilberto Barbosa Domont.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2021.

1. Aprendizado de máquina. 2. Word2vec. 3. Proteína. 4. Inteligência computacional. 5. Bioinformática. I. Menasche, Daniel Sadoc, orient. II. Domont, Gilberto Barbosa, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

ELVIS NOBREGA DE ALCANTARA

Classificação e aprendizado com proteínas de melanoma

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 4 de março de 2021

BANCA EXAMINADORA:

Daniel Sadoc Menasche
Prof. Daniel Sadoc Menasche

Participação por videoconferência
Prof. Gilberto Barbosa Domont

Participação por videoconferência
Prof. Fabio Cesar Sousa Nogueira

Participação por videoconferência
Dr. Bernard Kac

Participação por videoconferência
Prof. João Antonio Recio da Paixão

AGRADECIMENTOS

Primeiro gostaria de agradecer à minha família pelo apoio desde o início, meus pais por me criarem e minhas irmãs por me inspirarem tanto.

Também tenho a agradecer ao meu orientador por me acolher, obrigado também a todos os meus professores da UFRJ e da UNB, que com paciência construíram meu caminho acadêmico.

Um agradecimento especial ao Guilherme, que me apresentou o problema e tem trabalhado comigo. Sem você não teria feito esse trabalho, amigo.

Finalmente eu agradeço às pessoas que me transformaram profissionalmente nesse período da faculdade, todos os meus amigos do SIGA, do GRIS, do LabDIS e os Piratas do Cerrado. Muito obrigado a todos.

"Ours is the most exciting endeavor in biology.
Proteins provide the verbs to biology.
Think of an action in biology and you will find a protein behind it.
Think of a disease or its therapy and you'll find proteins there too."

Josh LaBaer

RESUMO

Exploração de metodologias diferentes de aprendizado de máquina para descobrir informações sobre o melanoma através de dados de pacientes. Foram usadas árvores de decisão nos dados clínicos a fim de montar uma hierarquia de dados sobre o paciente avaliados com sua sobrevida. As mesmas árvores também foram utilizadas nos proteomas dos pacientes com a intenção de classificar sua importância na possibilidade de metástase. Também foi executado um treinamento de rede neural a fim de encontrar representantes vetoriais das proteínas contidas nos proteomas dos pacientes, classificando por avaliação das informações das proteínas. O melanoma esconde tecnologias complexas para seu desenvolvimento e evasão, este trabalho apresenta algumas estruturas para repensar os dados que são conhecidos e que podem servir para uma nova interpretação da doença.

Palavras-chave: aprendizado de máquina. word2vec. proteína. inteligência computacional. bioinformática. árvore de decisão. melanoma. câncer. nlp. functionalprot2vec.

ABSTRACT

Exploration of several methodologies of machine learning to unveil informations about melanoma through patient data. Decision tree classifiers were used on clinical data in intent to architect a hierarchy of data about the survival rating of patients. The same trees were applied on patient proteomes in order to classify the significance in the probability of metastasis. Also was executed a training of a neural network in order to find embedded vectors of the proteins of the proteomes of patients, classified by evaluation of protein information. The melanoma hides complex technologies to his development and evasion, this work shows some structures to rethink the data that are known and can be used to make a new interpretation of the disease.

Keywords: machine learning. word2vec. protein. computational intelligence. bioinformatics. decision tree. melanoma. cancer. nlp. functionalprot2vec.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

UniProt	Universal Protein Resource
GF	fator de crescimento
pRb	proteína do retinoblastoma
CAM	moléculas de adesão célula a célula
VAE	autoencoders variacionais
DTC	Classificadores por árvores de decisão

SUMÁRIO

1 INTRODUÇÃO

De acordo com a OMS (Organização Mundial da Saúde), até os 74 anos de idade, a chance de desenvolver câncer é de 20.2% e em questão de mortalidade, o câncer é a segunda doença mais mortal, depois da isquemia cardíaca (MATTIUZZI; LIPPI, 2019). Existe uma crescente na prevalência e na mortalidade dessas doenças, o que destaca uma necessidade de novas tecnologias que revolucionem seu enfrentamento.

Dentre os cânceres, os de pele são vistos como manchas, ou feridas, e se desenvolvido através de melanócitos denomina-se melanoma. Os melanomas representam 4% de todos os tumores de pele, mas ainda assim marcam mais de 79% das mortes por câncer de pele.

Dada a colaboração entre a Unidade Proteômica e o Departamento de Ciências Clínicas da Lunds University, vários trabalhos foram feitos, uma deles iniciado em (BETANCOURT et al., 2019) que possibilitou o desenvolvimento nesse documento. Neste foram colhidos dados de 111 pacientes, clínicos e histopatológicos, e principalmente analisadas amostras dos tumores através de espectrometria de massas em tandem, um método capaz de medir a presença de proteínas. Essa análise possibilitou a quantização de 4963 proteínas, que depois foram filtradas através de um método de representatividade.

Existe larga pesquisa de descrição e anotação de dados sobre as proteínas, que ficam dispostos em variados bancos de dados, um exemplo deles é o UniProt. Essas anotações possuem descrições sobre localização, interação, até mesmo vias metabólicas as quais pertencem. Enriquecer os dados de presença de proteínas em pacientes com essas informações pode ser uma ferramenta valiosa na implementação de novos métodos.

A avaliação de novos métodos foi feita em duas vias: uma sobre interação direta com grafos, feita pelo colega Guilherme de Araujo Juvenal, orientado pelo professor Gilberto Barbosa Domont; e a desse documento que utiliza dos dados para entrelaçar contextos e gerar pontos através dessas anotações.

O objetivo do trabalho é criar e experimentar novas metodologias para avaliar o processo de desenvolvimento do melanoma. Seja usando grafos, redes neurais ou árvores, tentou-se estabelecer uma forma de retirar informação dos dados proteicos, como descritas a seguir:

- **Novo *embedding* de aspectos funcionais de proteínas:** propomos um novo *embedding* de aspectos funcionais de proteínas, que chamamos de functionalprot2vec. Enquanto os *embeddings* já existentes, como prot2vec, fazem uso apenas de informações estruturais das proteínas, nosso *embedding* foca em dados funcionais destas (e.g. angiogênese, localização, e outras funções)

- **Classificação automática de proteínas envolvidas em angiogênese:** usando árvores de decisão e o novo *embedding* aplicado anteriormente, classificamos de forma

interpretável as proteínas, identificado aquelas que tipicamente estão envolvidas em angiogênese

- **Classificação automática de fatores envolvidos na sobrevivência:** usando dados clínicos aliados às árvores de decisão, classificamos de forma interpretável fatores colhidos sobre pacientes, identificando aqueles que tipicamente estão envolvidos em maior sobrevivência dos mesmos

Conjuntamente, acreditamos que as duas contribuições acima aumentam nosso entendimento sobre a relevância de diferentes elementos envolvidos no câncer de pele do tipo melanoma, e abrem uma série de perguntas que dão margem para trabalhos futuros.

2 MELANOMA

O melanoma se caracteriza como uma pinta, ou mancha, na pele e é um câncer que se origina nos melanócitos, células responsáveis pela produção da melanina. Este costuma aparecer nas partes mais expostas ao sol. Apesar da maioria ter coloração marrom ou preta, alguns não são pigmentados.

Embora o melanoma seja parte de apenas 4% dos tumores de pele, ele representa mais de 79% das mortes por câncer de pele e nos países em desenvolvimento, a sobrevida média em cinco anos para essa doença é de apenas 56% (DIMATOS, 2009).

Apesar disso, o diagnóstico precoce e o tratamento cirúrgico são importantes fatores na cura dessa doença. Cerca de 90% dos melanomas são diagnosticados como tumores primários, sem a evidência de que o câncer se espalhou para outras partes do corpo, processo conhecido como metástase (DIMATOS, 2009).

2.1 BRESLOW

A classificação de melanomas é um desafio em si, dado que esse câncer é uma lesão muito imprevisível. O seu desenvolvimento é uma função de várias variáveis, das quais uma delas é o tamanho do tumor. Aproximadamente, a relação entre o diâmetro da lesão e a sobrevivência é inversamente proporcional (BRESLOW, 1970).

Desde 1950, muitos estudos objetivavam correlacionar o desenvolvimento da doença com a profundidade da invasão do melanoma na pele. Em 1969, Clark et al.[1] desenvolveram um método de classificação baseado nos níveis de microinvasão nas camadas da derme, camada abaixo da epiderme. Já em 1970, Breslow[2] introduziu um método para medir a profundidade da invasão em milímetros, tornando o sistema mais reproduzível e que tem mais correlação com a sobrevivência dos pacientes (MORTON et al., 1993).

Segundo Morton et al. (MORTON et al., 1993), ambos os métodos de Breslow e Clark tem uma grande significância no prognóstico quando avaliados por uma análise de múltiplos fatores, mas a espessura de Breslow é o fator mais efetivo porque ele é significativo mesmo em todos os níveis de Clark.

Essa medida que foi usada no trabalho pra classificar os pacientes que tiveram seus dados coletados. Os valores mudam de acordo com as organizações de médicos regionais, mas as que foram usadas como sendo o padrão de classificação são as dispostas na tabela 1:

---

[1] CLARK WH Jr., FROM L, BERNARDINO EA, MIHM MC **The histogenesis and biologic behavior of primary human malignant melanoma of the skin.**. Cancer Res 1969; 29:705-26

[2] BRESLOW A. **Thickness, cross-sectional areas and depth of inva- sion in the prognosis of cutaneous melanoma.**. Ann Surg 1970; 172:902-8

Tabela 1 – CLASSIFICAÇÃO DAS CATEGORIAS DE T

CATEGORIA T	ESPESSURA
T1	$\leq 1,0$ mm
T2	$> 1,0 - 2,0$ mm
T3	$> 2,0 - 4,0$ mm
T4	$> 4,0$ mm

Fonte: (E. LANDMAN G., 2017)

3 CONJUNTO DE DADOS

A fim de desenvolver o trabalho foram utilizados dois principais conjuntos de dados: dados de pacientes doentes e de pacientes saudáveis.

3.1 PACIENTES DOENTES

Os principais dados foram retiradas de um proteoma fornecido por uma parceria com o Departamento de Ciências Clínicas da Lund University, que foi depositado no banco de dados ProteomeXchange[1] com a identificação PXD009630[2].

Proteoma principalmente composto de melanócitos que infiltraram os linfonodos. Neste, 11802 proteínas foram identificadas, tendo um terço delas presente em mais da metade das amostras.

Os dados foram retirados de 111 pacientes com melanoma que foram analisados de 1975 a 2011. Num total de 68 homens e 43 mulheres, de uma média de 62,4 e desvio padrão de 13,7 anos de idade.

Também são providos dados histopatológicos dos pacientes, como a espessura Breslow e índice Clark.

Tabela 2 – ESCALA BRESLOW

LARGURA BRESLOW	NÚMERO	PORCENTAGEM
<1	11	10%
<2	26	23%
<3	23	21%
<4	27	24%

3.2 PACIENTES SAUDÁVEIS

Com a finalidade de estabelecer uma base para a comparação entre as classes dos pacientes doentes, foi utilizado um conjunto de dados extraído do banco de dados The Human Protein Atlas[3]. Este contém o transcriptoma da pele humana saudável, que gera um total de 14857 proteínas. Os dados pode ser divididos de acordo com os níveis de expressão do RNAm da pele em comparação com os demais tecidos, da seguinte forma: 547 proteínas de elevada expressão na pele, 5985 proteínas de elevada expressão na pele e outros tecidos e 8325 expressos em quase todos os tecidos, mas com baixa especificidade.

[1] http://www.proteomexchange.org
[2] (NUNEZ, 2019)
[3] https://www.proteinatlas.org/humanproteome/tissue/skin

3.3 UNIPROT

Além dos dados de pacientes também foram usadas anotações sobre as proteínas, que indicam informações como quais proteínas interagem entre si, quais seus nomes, suas localizações, etc. Essas anotações foram recuperadas do banco de recursos de proteínas UniProt[4].

O UniProt é uma colaboração entre o European Bioinformatics Institute (EMBL-EBI)[5], o SIB Swiss Institute of Bioinformatics[6] e o Protein Information Resource (PIR)[7]. Este site é constituído de três bancos de dados, o UniProtKB, o UniRef e o UniParc. O UniProtKB é a principal base de informações sobre proteínas, o UniRef contém grupos de informações retiradas do UniProtKB, O UniParc é um banco de apoio com descrição das proteínas a fim de identificar elas em outros bancos de dados externos.

O UniProtKB é o que foi usado no trabalho. Ele é constituído por duas partes: o Swiss-Prot e o TrEMBL. O Swiss-Prot é constituído de informações manualmente anotadas, enquanto que o TrEMBL é gerado a partir de informações de bancos variados. Para o trabalho, foram usadas anotações do Swiss-Prot, mais especificamente as seguintes:

- **Entrada:** código constituído por letras e números que identifica a proteína no UniProt, ela é única para cada proteína. Se uma entrada for separada ou unida, o código que a descreve não é perdido, é guardado como secundário, enquanto sempre se tem um código primário que é único para cada proteína.

- **Nome da proteína:** provém uma lista exaustiva com todos os nomes que a proteína pode ter, incluindo obsoletos. Além disso possui uma breve descrição das atividades desta.

- **Localização subcelular:** informações sobre localização e topologia de uma proteína madura no interior da célula. Toda a ontologia é descrita, se conhecida. Quando diferentes locais coexistem são separados por ponto final.

- **Interage com:** aponta informação binária de interação com outras proteínas do banco, expostas em ordem alfanumérica, exceto com interação homóloga, onde é citada ao início, com a palava "itself".

- **Função:** descreve em texto as funções gerais da proteína.

- **Polimorfismo**: possui uma definição das diferentes formas que a proteína pode ter, baseada numa análise comparativa das sequências de aminoácidos sem ordem.

[4] https://www.uniprot.org
[5] https://www.ebi.ac.uk/
[6] https://www.sib.swiss/
[7] http://pir.georgetown.edu/

- **Domínio:** descreve a posição e tipo de domínio. Sendo domínio caracterizado como uma região da cadeia de aminoácidos que formam um conjunto especifica de estruturas secundarias em três dimensões.

- **Ontologia gênica:** Ontologia gênica refere-se ao produto de um determinado gene e à função que ele desempenha na maquinaria celular a nível molecular, se refere ao banco de dados externo Gene Ontology[8]. Deste foram usados três níveis hierárquicos:

    1. **Componente celular:** descreve a localização celular na qual a proteína é ativa.
    2. **Processo biológico:** que se refere à série de eventos realizados por uma ou mais funções celulares (via metabólica). Por exemplo: cadeia respiratória, glicólise.
    3. **Função molecular:** a atividade que uma proteína desempenha no meio celular. Por exemplo: formação de filamento, agente transportador de elétrons.

[8] http://geneontology.org/

4 REPRESENTAÇÃO DO CONHECIMENTO (*EMBEDDING*)

O word2vec é uma metodologia criada em (MIKOLOV et al., 2013), composta por dois modelos de arquiteturas de redes neurais que objetivam computar representações vetoriais contínuas de palavras, para grandes conjuntos de dados. São redes que, através de uma medida de similaridade geram pontos para cada diferente palavra num texto.

4.1 ARQUITETURAS

Muitos modelos já haviam sido propostas para estimar uma distribuição contínua de palavras, dentre elas a Latent Semantic Analysis (LSA) e a Latent Dirichlet Allocation (LDA). A principal vantagem do word2vec para estas representações é seu menor custo computacional.

Para todos os modelos word2vec, tem-se a seguinte complexidade:

$$O = E \times T \times Q$$

Onde E é o número de épocas de treino, T o número de palavras no conjunto de treino e o Q é o valor único para cada uma das implementações do modelo. O valor costumeiro para $E = 3 \rightarrow 50$ e T de até um bilhão.

4.1.1 Representação das palavras

Inicialmente as palavras são separadas num índice de vocabulários, cada sequência de letras que representam uma palavra são convertidas para um valor. Essa representação é obtida através do método conhecido como one-hot encoding. Neste método, a dimensão dos vetores é igual à quantidade de palavras do vocabulário e cada palavra tem sua dimensão representante, ou seja, cada palavra tem como valor um na sua dimensão e as outras são zeradas. A figura 1 mostra um exemplo de três palavras num contexto que existem oito palavras, elas aparecem como responsáveis pelas dimensões dois, cinco e sete.

4.1.2 Continuous Bag-of-Words (CBOW)

Nessa arquitetura a camada de projeção resume as palavras do mesmo contexto, todas as palavras são projetadas na mesma posição. Essa arquitetura é chama de bag-of-words, "bolsa de palavras", numa tradução livre, porque a ordem das palavras não influencia na projeção.

A melhor eficiência do experimento de (MIKOLOV et al., 2013) foi encontrada com um contexto de quatro palavras anteriores e quatro posteriores e o objetivo é classificar a palavra do meio.

Figura 1 – Exemplo de One-Hot Encoding

Sua complexidade de treino é dada por:

$$Q = N \times D + D \times \log_2(V)$$

N são o número de palavras, V o número de palavras no vocabulário e D a dimensão escolhida para a camada escondida.

Esse modelo é diferente do bag-of-words anterior porque ele usa uma distribuição contínua das palavras no contexto. Sua arquitetura é mostrada a seguir:

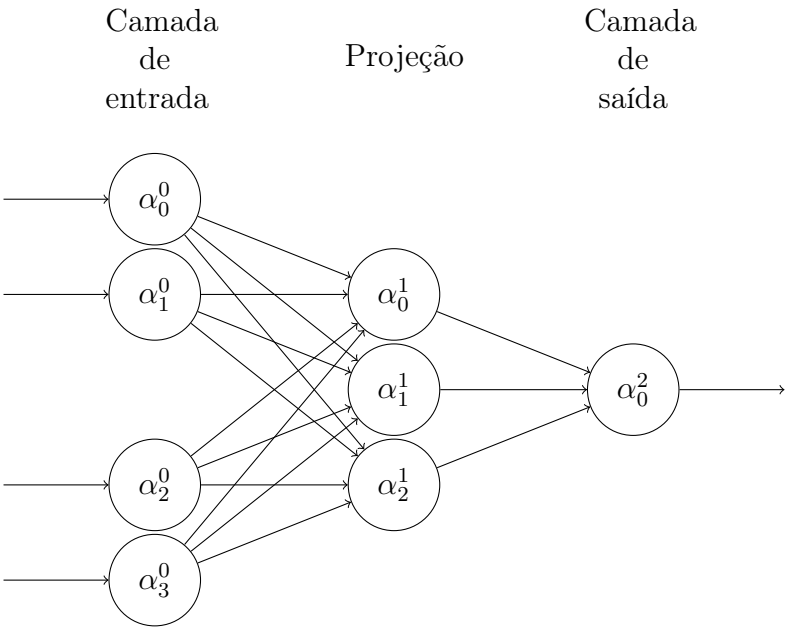

Figura 2 – Continuous Bag of Words

A primeira camada, de entrada, tem em cada neurônio uma palavra, sendo α_0^0 e α_1^0 as palavras que vem antes, no contexto de treino, e α_2^0 e α_3^0 as que vem depois. No

exemplo a projeção tem três dimensões, mas esse valor (D) pode ser definido com o número preferível, o que vai definir a dimensão do espaço latente das palavras. A camada de saída é composta pela palavra central do contexto, a que deverá ser "prevista"pela rede, como objetivo do treino.

4.1.3 Continuous Skip-gram

Essa arquitetura é muito similar à anterior, mas o objetivo é invertido. Ao invés de predizer a atual palavra usando o contexto, ela tenta maximizar a classificação de uma palavra para as outras da mesma sentença. Ela usa a palavra atual como entrada num classificador log-linear de uma camada de projeção e prevê as palavras do contexto que fazem parte de um alcance anterior e posterior à atual.

Sua complexidade é regida por:

$$Q = C \times (D + D \times \log_2(V))$$

Tendo C a distância máxima entre as palavras, ou seja, se uma palavra for escolhida haverá até C palavras anteriores e posteriores como rótulos de saída adequados. O CBOW tenta prever a palavra de acordo com o contexto, enquanto que o Skip-gram tenta dizer o contexto a partir da palavra índice.

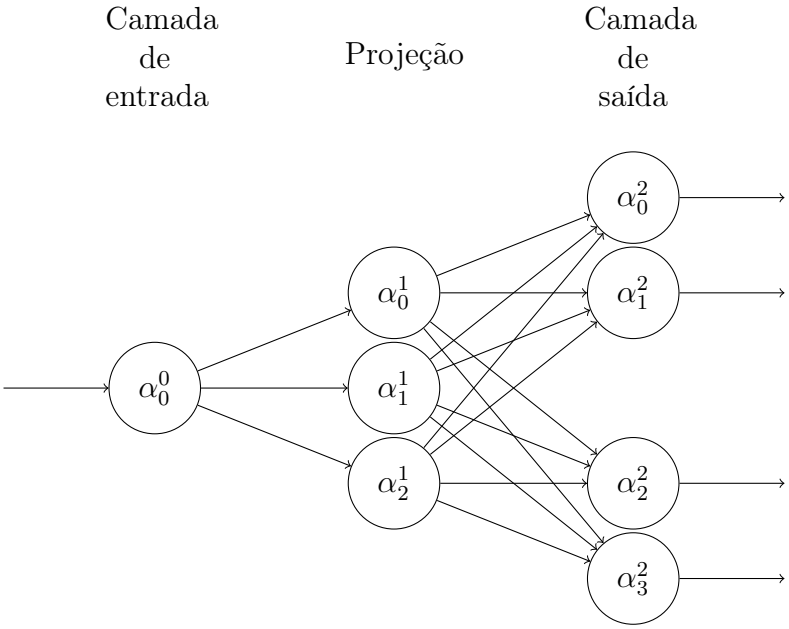

Figura 3 – Continuous Skip-gram

Pegue, por exemplo, um texto como "Só se vê com o coração, o essencial é invisível aos olhos". Primeiro são retiradas as conjunções e artigos, as *stop words*, tendo assim "Só vê coração essencial invisível olhos"para o treinamento, que teria a seguinte sequência:

- **Só** vê coração
- Só **vê** coração essencial
- Só vê **coração** essencial invisível
- vê coração **essencial** invisível olhos
- coração essencial **invisível** olhos
- essencial invisível **olhos**

5 ÁRVORES DE DECISÃO

Mais uma ferramenta que dá suporte ao processo de recuperação das informações são os classificadores com árvores de decisão (DTC). Elas são um tipo de classificador que cria uma árvore com probabilidades sobre uma hipótese, a cada passo é avaliada uma característica dos dados, assim chegando na probabilidade da resposta.

Tendo (X, Y) variáveis aleatórias juntamente distribuídas. X é um vetor de dimensão q, se os valores de interesse forem números reais tem-se X distribuído em R^q. Já Y toma valores inteiros $\{1, 2, 3...J\}$, se houverem J classes. O objetivo de todo classificador, e da árvore de decisão em especial, é estimar Y dado um valor observado de X (SAFAVIAN; LANDGREBE, 1991).

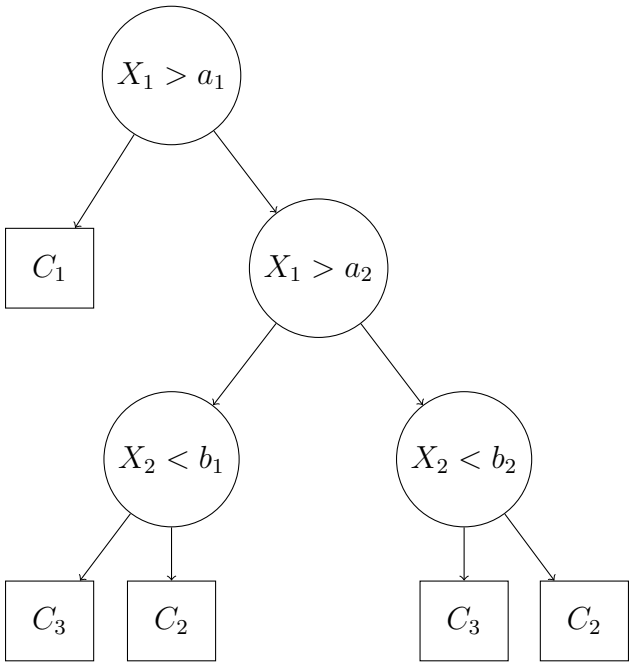

Figura 4 – Árvore de decisões simples

Na figura 4 tem-se um exemplo que ilustra uma árvore de decisões, os nós desenhados com círculo denotam as regras de decisão modeladas. Os nós terminais da árvore, retângulos, representam as classes que são decididas através da execução do método dada uma entrada X.

Essas árvores podem ser montadas com vários métodos diferentes, o conceito principal é tentar definir quais são as características que possuem mais entropia, que separam melhor cada uma das linhas de dados. Dessa forma, haverá nos primeiros nós da árvore as características que carregam mais informação sobre a separação das classes, denotando uma importância dessa característica.

Exemplos de algoritmos de DTC são o ID3 e o CART (PRIYAM et al., 2013):

- O ID3 (Dicotomizador iterativo 3) constrói a árvore de decisão empregando uma pesquisa gulosa, *top-down* através dos conjuntos para testar os atributos a cada nó. Para escolher a característica mais importante para classificar um conjunto, usa-se uma métrica conhecida por ganho de informação. Assim, essa função define o quão balanceado é a aplicação do ID3, mas é sensível a ruído já que as avaliações são feitas de forma gulosa.

- O CART foi implementado para ser capaz de construir árvores de classificação e regressão. A montagem da árvore é baseada em separações binárias dos atributos, e usa a medida índice de Gini para selecionar cada um deles. Ele é único por usar análise de regressão que pode ser usada para prever variáveis dependentes dado um conjunto de variáveis preditoras durante um período de tempo.

6 MÉTODO

Os pacientes dos quais tiveram suas proteínas detectadas foram separados em estágios, através de sua classe Breslow, tendo assim quatro classes de interesse (BC1, BC2, BC3 e BC4), além da classe saudável (HH). O principal foco é usar suas proteínas para treinar uma rede neural diferente para cada uma das classes, para depois comparar os valores resultantes para cada classe. Esses dados também foram utilizados nos métodos baseados em árvores de decisão, nos quais tiveram proteínas classificadas pelas suas funções.

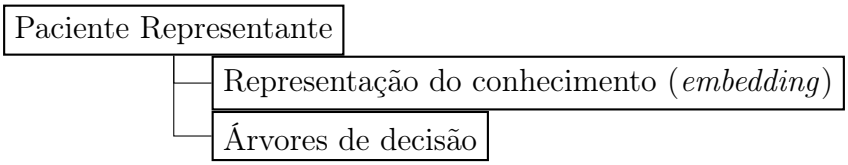

Figura 5 – Fluxo de dados

Uma representação em fluxograma dos passos seguidos pelo desenvolvimento podem ser vistos na figura 6:

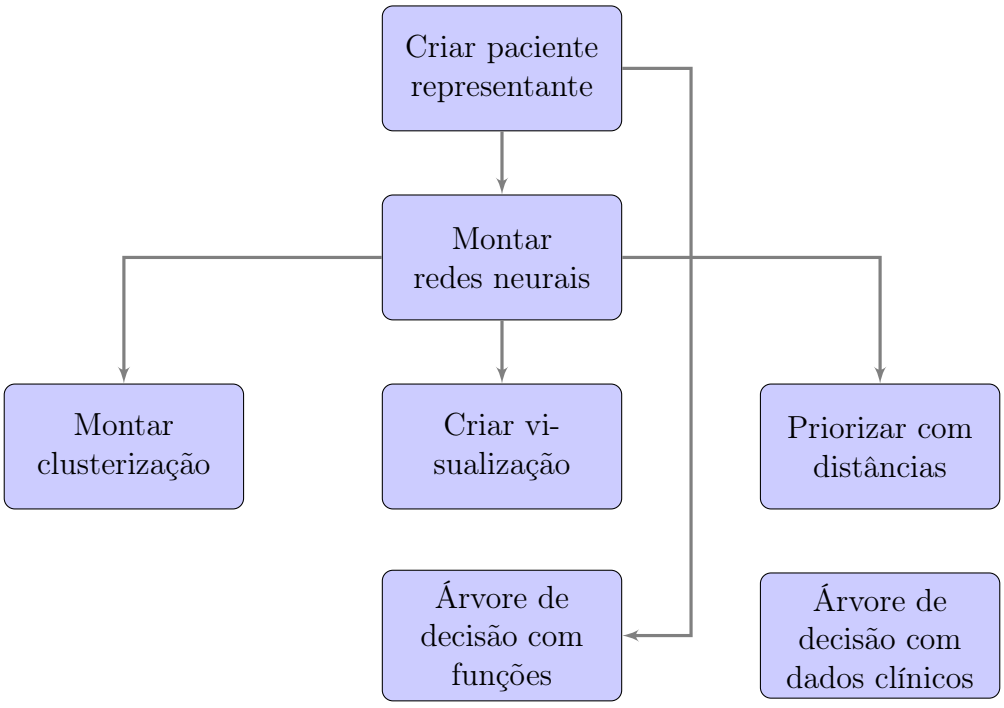

Figura 6 – Fluxograma de desenvolvimento

Depois de montar os bancos de dados com pacientes representantes foram executadas as duas estratégias usando seus dados, a representação do conhecimento (*embedding*) e a de árvores de decisão. As redes neurais disponibilizaram os pontos que foram aplicados em vários métodos diferentes, clusterização, visualização e priorização. Já as árvores usaram

o paciente representante na aplicação de funções biológicas, mas no caso da dados clínicos foi-se pensada uma base de dados diferente.

6.1 PACIENTE REPRESENTANTE

Cada proteoma de um paciente possui todas as proteínas encontradas no linfoma doente, ou na pele, no caso do proteoma saudável. A primeira decisão relacionada aos dados era como criar um paciente representante para que esse fosse usado no modelo proposto. Existem três alternativas:

1. Utilizar todas as proteínas que foram identificadas em todos os proteomas. Esse método utiliza todos os dados coletados em laboratório, mas por outro lado inclui todas as proteínas que possivelmente não estão relacionadas ao câncer, mas à alguma atividade única do paciente. Este seria como utilizar uma união de todas as proteínas detectadas em pacientes.

2. Outra estratégia é utilizar apenas as proteínas que foram identificadas exclusivamente em todos os proteomas. Esse método é bem mais restrito, ele remove as possíveis proteínas únicas dos pacientes, porém também ignora as proteínas que estão na maioria dos conjuntos. Este seria o caso de uma interseção das proteínas.

3. A última forma de se obter um paciente representante é através de algum método de representatividade que estabelecesse um limiar que definisse a inclusão da proteína ao conjunto representante. Este seria um meio termo entre os métodos de união e interseção.

O método utilizado no trabalho foi o intermediário utilizando o método de proporção populacional. Tal decisão foi tomada por existirem limitações na sensibilidade da técnica que monta os proteomas, sendo possível uma proteína não ser identificada mesmo existindo na amostra; também é possível que proteínas sejam específicas do paciente relacionadas à sua vida pessoal ou características específicas das suas células. Como o objetivo é descobrir informações sobre a natureza da doença, foi escolhido não adicionar partes específicas.

Como a população dos testes era constituída por cerca de 20 pacientes, foi usado o método de representatividade para pequenas populações descrito em (TRIOLA, 2005):

$$n = \frac{N \cdot \hat{p}\hat{q} \cdot [z_{\alpha/2}]^2}{\hat{p}\hat{q} \cdot [z_{\alpha/2}]^2 + (N-1) \cdot E^2}$$

Para uma proteína ser representativa ela deveria estar em n dentre todos os N proteomas. Os valores de $\hat{p}$ e $\hat{q}$ podem ser estimados em $0,5$ e significam a proporção populacional de amostras que pertencem ou não à categoria de interesse, respectivamente. Já $z_{\alpha/2}$ é o valor crítico correspondente ao grau de confiança, no caso usado $1,96$ do desvio padrão,

o bastante para um índice de 95% de confiança. O valor de E significa a margem de erro, ou erro máximo, definida no trabalho como sendo de 2%.

6.2 REPRESENTAÇÃO DO CONHECIMENTO (*EMBEDDING*): FUNCTIONALPROT2VEC

De início, o objetivo foi descobrir novas informações sobre funcionamento e interações relacionados ao melanoma. Para se ter uma nova visão do problema, planejou-se usar da transformação de palavras em vetores através do seu contexto, vinda pelo treino de uma rede neural, como mostrado no método *word2vec*.

Ao se pesquisar sobre metodologias já implementadas, em trabalhos relacionados, foram encontradas poucas aplicações, delas a mais notável foi a (BUCHAN; JONES, 2020) por ser a mais bem detalhada. O artigo descreve o uso de *word2vec* com sequências de proteínas e seus domínios como contexto, os testes foram feitos em relação à predição das famílias de proteínas. A similaridade de sequências biológicas com gramáticas é um método bastante usado cientificamente, como no exemplo do artigo mais recente (HIE et al., 2021), que usa o gene de vírus para tentar recuperar informações sobre eles.

Porém, na implementação proposta por esse trabalho difere dos antigos trabalhos que focam em sequenciamento no serviço de descobrir informações, ela define o contexto como sendo os nomes que descrevem a proteína no banco de dados UniProt. Assim, anotações como proteínas relacionadas são matéria do treinamento da rede em questão de relacionamentos, e vice-versa, as palavras também são treinadas entre si e com as próprias proteínas.

6.2.1 Geração de tokens

Processo inicial do tratamento dos dados, nele são retiradas as palavras que não possuem conteúdo semântico, que não ajudam no treinamento e também há uma normalização para que palavras idênticas tenham o mesmo valor. Ao seu fim são definidos "tokens"que representam as palavras.

Inicialmente foi utilizado o NLTK[1] ferramenta para a linguagem Python que auxilia no desenvolvimento com linguagem natural. Ela tem métodos para análise, classificação, geração de tokens e outros. No trabalho esta foi utilizada para remover *"stop words"* da língua inglesa, palavras como *and*, *the* e *is*. Depois desse passo tem-se apenas as palavras com conteúdo semântico.

Para as palavras restantes foi executada uma indexação, à cada palavra é atribuída um índice único, que vai representar sua dimensão no *one-hot encoding*, assim todas as palavras se tornam vetores esparsos de dimensão igual ao número de vocábulos de todo o dataset.

[1] https://www.nltk.org/

6.2.2 Criação dos contextos

No trabalho a definição de contexto compreende todas as anotações retiradas do banco UniProt para cada uma das proteínas encontradas no proteoma representante. O objetivo é relacionar todas as palavras que descrevem a proteína alvo, entre si e também com a proteína. Por exemplo, se uma proteína tiver como anotação a localização núcleo, núcleo e o nome da proteína estarão no mesmo contexto no treinamento. Todas essas informações entram como características da proteína, ela possui um identificador, mas ao mesmo tempo tem nomes e funções, seus dados podem ser encontrados através de todos estes elementos, alguns podem descrever melhor, outros não tão significativamente.

A rede tem um tamanho de entrada igual ao número de palavras do maior contexto para que todos os treinamentos sejam possíveis. Assim, toda vez que as palavras não atingirem o número igual ao do maior contexto houve a repetição de cada linha até o preenchimento da matriz.

6.2.3 Rede neural

As principais ferramentas utilizadas para se desenvolver a rede foram Keras[2] e TensorFlow[3]. O TensorFlow oferece uma plataforma com várias ferramentas diferentes que auxiliam no processo de configuração de redes neurais, tem suporte para o uso da placa gráfica no treinamento, além de tratamento de dados e até visualização. O Keras é responsável por oferecer uma API mais simplificada do processo de configurar a rede, tem um código mais descritivo enquanto que funciona muito bem em conjunto com o TensorFlow.

A rede criada tem a seguinte arquitetura:

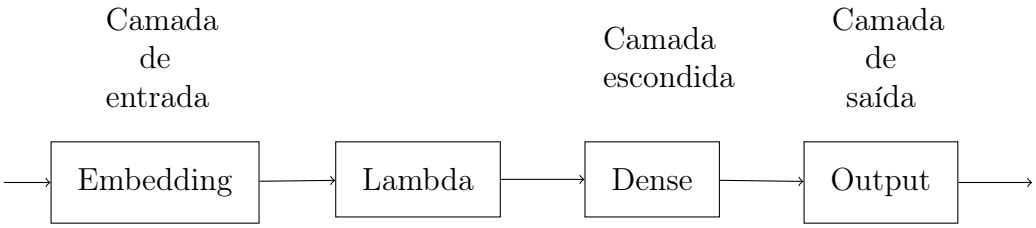

Figura 7 – CBOW utilizada

Três dessas camadas foram implementadas utilizando um padrão do Keras onde o usuário pode descrever uma sequência pela qual os dados passarão pela rede, as *layers*. Elas são *Embedding*, *Lambda* e *Dense*.

- A camada de *Embedding* guarda as informações de incorporação, o principal objetivo da rede. Ela vai ter o valor dos vetores, ao final do treino. De entrada ela tem o

[2] https://keras.io/
[3] https://www.tensorflow.org/

número de neurônios igual ao tamanho do maior contexto, e cada neurônio tem de entrada um vetor de tamanho igual ao número de vocábulos de todo o documento.

- Lambda é uma camada interna de auxílio, do Keras, que aplica uma função a todos os dados que passam por ela, no caso é executada uma média, como orientado por (MIKOLOV et al., 2013), ajudando a atingir resultados mais sólidos.

- A camada interna é completamente conectada, assim chamada de *Dense*. Ela tem de neurônios a dimensão desejada para a normalização dos dados, é interessante que tenha tamanho bem menor que a de entrada, para que os dados passem por uma redução de dimensionalidade.

- A camada de saída *Output* tem que ter o número de neurônios iguais aos da entrada, para que o treinamento seja capaz de não ser supervisionado, apenas usando os próprios valores da entrada como avaliação.

Depois de treinada a rede neural pode-se atingir os valores para cada vocábulo como sendo as conexões entre a primeira camada e a camada interna, cada conexão vai ter um índice que representa uma palavra diferente, por exemplo, todos os primeiros índices conterão o valor do primeiro termo. Sendo assim, cada palavra tem um vetor de tamanho igual ao número de neurônios da camada interna, no caso do trabalho é igual a cem.

Um exemplo de palavra resultante é mostrado no quadro 1, no caso do treinamento da segunda classe de Breslow, para uma rede com cem neurônios internos, de forma resumida.

Quadro 1 – RESULTADO DE TREINO DE UMA PALAVRA

TERMO	0	1	⋯	98	99
ANGIOGENIC	0.038502607	0.032154907	⋯	-0.000344444	-0.023104405

6.3 ÁRVORES DE DECISÃO

Enquanto se tentava formas de extrair informação dos pontos que resultaram do método de functionalprot2vec, foi discutida a possibilidade de aplicar árvores de decisão nos dados brutos. Essas árvores possuem aplicação direta em exemplos de bancos de dados textuais, ao usar o *one-hot encoding*, o que pode retornar informações interessantes com interpretabilidade mais simples do que o método anterior.

6.3.1 Árvores de funções biológicas

O primeiro dos modelos montados foi o de proteínas por funções biológicas. Nele foi feita a criação de árvores de decisão usando informações das proteínas sobre o pertenci-

mento em classes funcionais como registrado no banco de dados *Gene Ontology*[4], o maior banco de dados sobre funções.

Quadro 2 – EXEMPLO DE DADOS PARA ÁRVORE

NOMES	GO:A	GO:B	GO:C	GO:D	GO:E
Proteína 1	0	1	0	0	1
Proteína 2	0	1	1	1	0
Proteína 3	1	0	0	0	0

O quadro 2 acima demonstra um exemplo de como os dados foram organizados para o treinamento. Cada proteína tem suas anotações, de acordo com o *Gene Ontology*, de funções biológicas, representadas por um identificador GO, dois pontos e seu número, no exemplo substituídos por letras. Ter a função significa ter o valor binário positivo, igual a um, e não ter, zero. A lista de proteínas é retirada dos dados criados sobre pacientes representantes de cada classe Breslow, assim sendo possível verificar mudanças relacionadas ao desenvolvimento de tamanho físico.

Depois de organizar os dados em valores classificatórios, sobre o critério estratégico e variável de que função é a de interesse para o resultado. Foi definida a função de angiogênese, de identificador igual a "GO:0001525"e algumas relacionadas, como regulação de angiogênese: "GO:0045765", dado que o interesse principal da classificação seria a possibilidade do tumor de ser metastático e essa função é essencial para que a doença atinja esse nível de desenvolvimento.

O artigo do qual gerou a iniciativa da pesquisa e que é responsável pelos dados (BETANCOURT et al., 2019) também fez uma avaliação similar, mas sem o uso de DTCs e avaliando pela sobrevida. Os autores inicialmente fizeram um agrupamento com classificação não supervisionada baseada em consenso, então aplicaram uma metodologia supervisionada Partial Least Squares (PLS) em combinação com a Cox Proportional Hazards modeling (PLS-Cox). Ambas as tentativas produziram agrupamentos com diferenças significantes em sobrevivência.

O agrupamento não supervisionado produziu três agrupamentos com distintas classes de sobrevivência (BETANCOURT et al., 2019). Já com o método PLS-Cox separou 27 proteínas como sendo as mais importantes para a variável, e depois de agrupá-las retornou-se três agrupamentos, mesma quantidade do anterior.

6.3.2 Árvores de dados clínicos

Depois de montar as árvores de decisão relacionadas às proteínas e suas funções biológicas, pensou-se em dar um passo atrás, tentar desenvolver resultados de dados mais simples e menos específicos, como os dados clínicos de todos os pacientes. No banco há

[4] <http://geneontology.org/>

dados de classes como sexo biológico, idade, tempo de sobrevida, até mesmo em que partes do corpo a doença se espalhou. Dessa forma se construiu também uma árvore com os dados clínicos para cada paciente, e de dado objetivo escolheu-se a sobrevivência do paciente, valor binário.

O artigo (BETANCOURT et al., 2019) também fez avaliação dos dados clínicos junto dos histopatológicos (relacionados à medidas microscópicas da doença). Foi utilizada uma análise não supervisionada multivariada através da análise de componente principal (PCA) e não houve muita separação representativa, exceto pelo conteúdo de célula tumoral. O quão heterogêneo o conteúdo da célula estava determina variabilidade nos dados proteômicos.

Os dados utilizados no método foram:

- **gender**: o gênero do paciente
- **stage**: fase da metástase, valor numérico (1=local, 2=em trânsito, 3=regional, 4=distante)
- **age**: a idade do paciente
- **Bres.class**: a classe Breslow a qual o paciente se encaixa
- **prim.trunk, prim.head/neck, prim.extremity.upper, prim.extremity.lower, prim.other**: valor binário que determina onde o câncer começou, onde houve a primeira ferida (trunk=tronco, head/neck=cabeça/pescoço, extremity.upper=extremidades superiores, extremity.lower=extremidades inferiores, other=outras regiões)
- **spread.trunk, spread.head/neck, spread.extremity.upper, spread.extremity.lower, spread.other**: valor binário que determina para onde a doença se espalhou, alvos de metástase (trunk=tronco, head/neck=cabeça/pescoço, extremity.upper=extremidades superiores, extremity.lower=extremidades inferiores, other=outras regiões)
- **BRAF**: existência do oncogene homônimo, valor binário
- **NRAS**: existência do oncogene homônimo, valor binário
- **dead by melanoma**: sobrevivência do paciente, valor binário

7 RESULTADOS

Após serem encontrados os pontos através das palavras, a primeira estratégia que foi pensada no objetivo de retirar informações é a de agrupamento. Se trata de tentar atribuir grupos para os pontos, para depois atribuir anotações, além de ser possível uma tentativa de classificação de importância, por exemplo por quem se destaca geometricamente nesses grupos. Como exemplo tem-se a visualização da figura 8, que destaca todas as proteínas iniciadas em 'P' na nuvem de pontos criada do paciente representante na segunda classe *Breslow*.

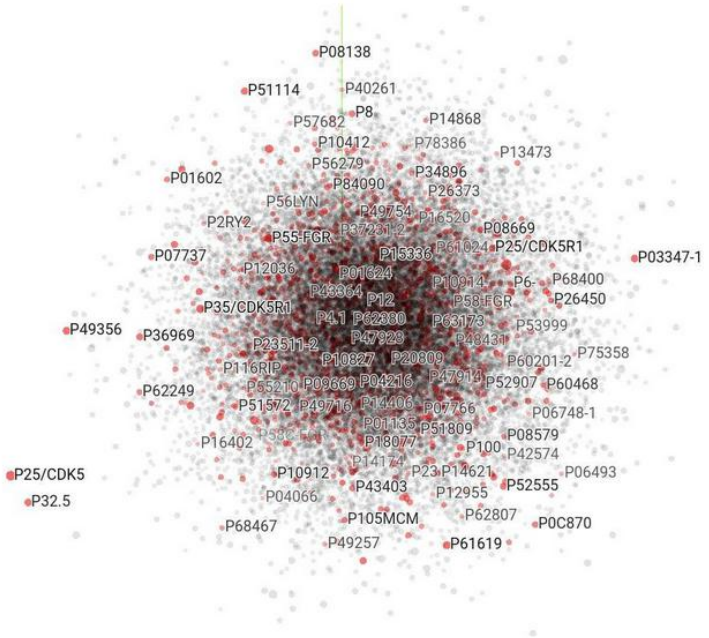

Figura 8 – Visão simplificada dos pontos das proteínas

7.1 REPRESENTAÇÃO DO CONHECIMENTO (*EMBEDDING*): FUNCTIONALPROT2VEC

Inicialmente houve a tentativa de executar um agrupamento dos pontos. Utilizou-se alguns métodos que verificava-se a existência de clusters e alguns que separavam estes. Mas os resultados não foram tão interessantes, já que não se passava de um cluster, com no máximo pontos distantes dele, por isso foi feita a pergunta de que pontos seriam estes, e isso poderia aparecer num método de visualização.

Para a visualização foi utilizada a ferramenta do Tensorflow disponível online[1]. Nela existem vários alguns métodos que podem ser utilizados para possibilitar visualizar os pontos que podem ser de milhares de dimensões, dentre eles tem o PCA (análise do

[1] <https://projector.tensorflow.org/>

componente principal) e o T-SNE (conexão estocástica usando t-student entre vizinhos). Foi aplicada o último método nos dados do trabalho, como visto na figura 9.

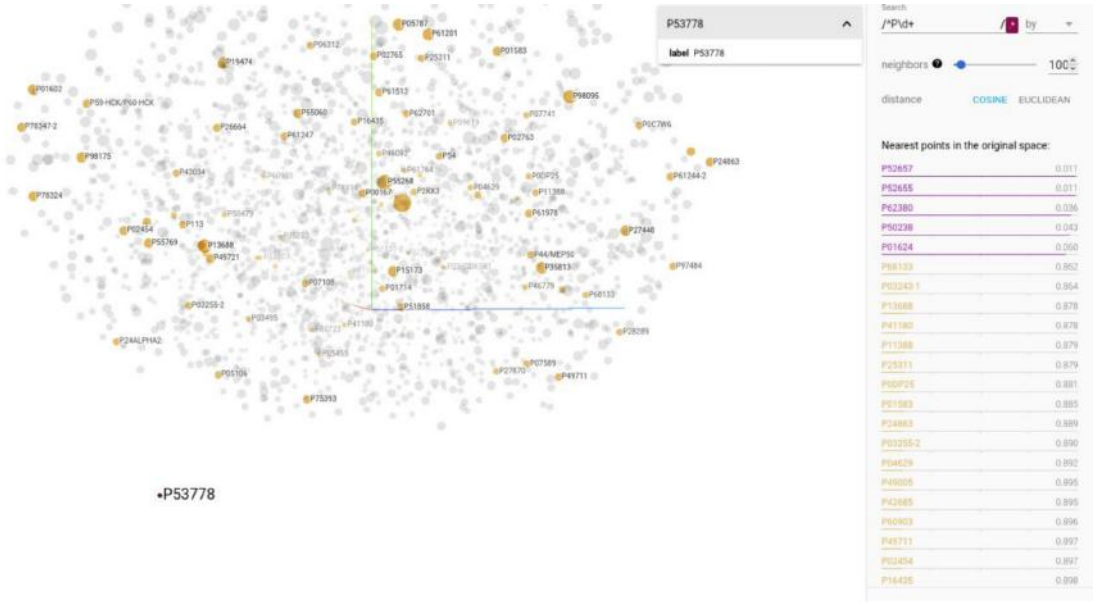

Figura 9 – Visão T-SNE das proteínas BCr2

Foi verificado, para a segunda classe *Breslow*, que algumas proteínas se destacavam do grupo principal, se distanciando, como a P53778 vista na figura 9. Isso significa que em algumas das três dimensões ela tem valores estocasticamente maiores em relação com os outros pontos. A proteína P53778 é uma proteína quinase (enzima que modifica fosfatos de proteínas) ativada por mitógeno (substância que induz mitose, multiplicação celular), conhecida por bloquear o ciclo celular como resposta a danos de DNA (WU et al., 2010), impedindo que a célula alvo se multiplique. E nas mais próximas geometricamente dela no espaço original destaca-se a P50238, proteína rica em cisteína (aminoácido) é uma proteína que promove invasão e migração celular, como visto nos artigos (HE et al., 2017) e (ZHANG et al., 2018).

Outra forma pensada foi a de distâncias dos pontos a nomes conhecidos. Por exemplo, pode ser aplicado um produto interno entre todos os pontos e o representante do termo angiogênese. Nesse exemplo foram ordenados os maiores valores e filtradas apenas as palavras representantes de proteínas, assim tendo como os resultados de maior valor os seguintes:

- P27482 e Q56UQ5, proteínas ligantes de cálcio, responsáveis por sinalização de equilíbrio bioquímico intracelular. Essas proteínas não são anotadas como relacionadas a câncer e nem angiogênese, mas foi descoberto, como visto em (MUNARON, 2012) e (MOCCIA et al., 2019), que sinalização de cálcio está relacionada a aumentada vascularização e também multiplicação celular.

- Q9BRA2 responsável pela peroxidase, já anotada como relacionada a sinalização de uma via metabólica de necrose de tumores, uma resposta das células a tumores. De significância em cânceres como visto em (CHO et al., 2019) e outros artigos.

7.2 ÁRVORES DE DECISÃO

7.2.1 Árvores de funções proteicas

Ao ser aplicado o treinamento com DTC para os dados funcionais das proteínas obteve-se a árvore mostrada na figura 10. Nela se destacaram como mais importantes as funções de regulação negativa de fibrinólise e de diferenciação de queratinócito.

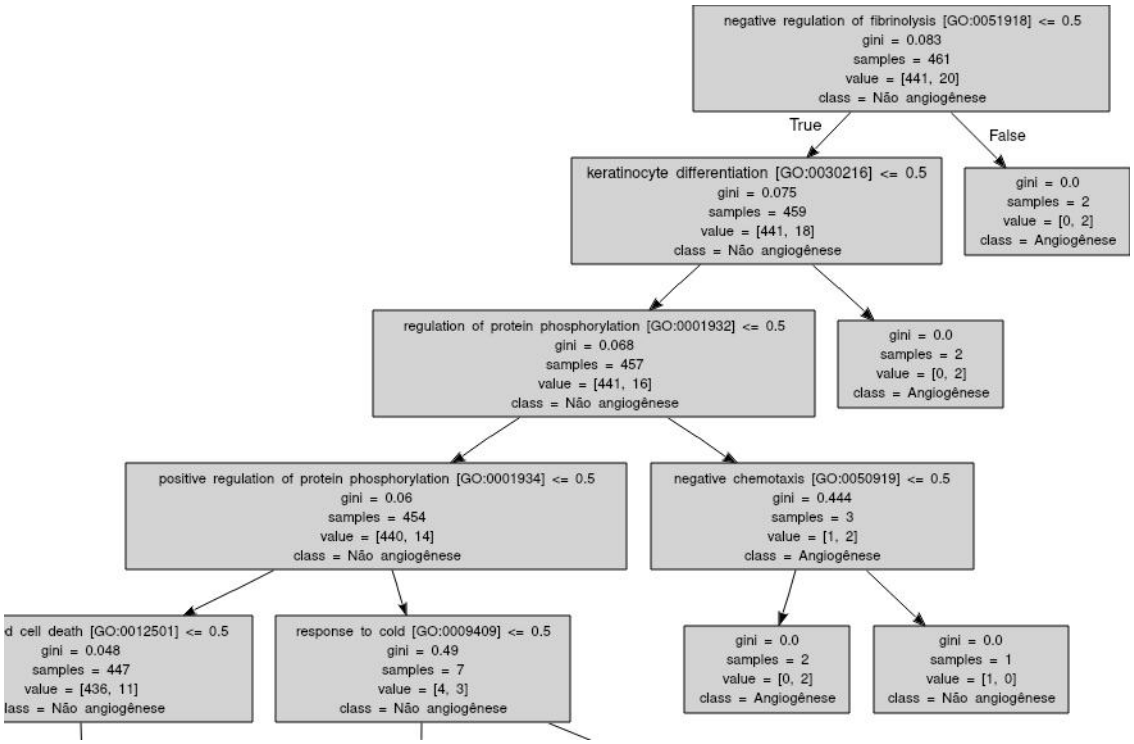

Figura 10 – Árvore de funções relacionadas à angiogênese

- **GO:0051918 - Regulação negativa de fibrinólise**: processos que previnem, param ou diminuem a taxa da fibrinólise, destruição de coágulos de fibrina. A fibrina é uma proteína responsável por coagulação, que é inibida pela trombomodulina, outra proteína que quando em baixa expressão tem histórico de menores chances de metástase (HOSAKA et al., 2000).
- **GO:0030216 - Diferenciação de queratinócito**: Queratinócitos são células da pele responsáveis pela síntese de queratina, em homeostase elas controlam o crescimento e comportamento dos melanócitos, células produtoras de melanina. Se esse equilíbrio for quebrado, tumores podem surgir (HAASS; SMALLEY; HERLYN, 2004).

Também foi montada uma árvore para um modelo onde proteínas que são relacionadas à proteínas anotadas como tendo funções de angiogênese também são marcadas como proteínas de interesse. Assim, se uma proteína interage com uma proteína marcada no método anterior ela também seria marcada como alvo. Através desse modelo foi montada a árvore da figura 11

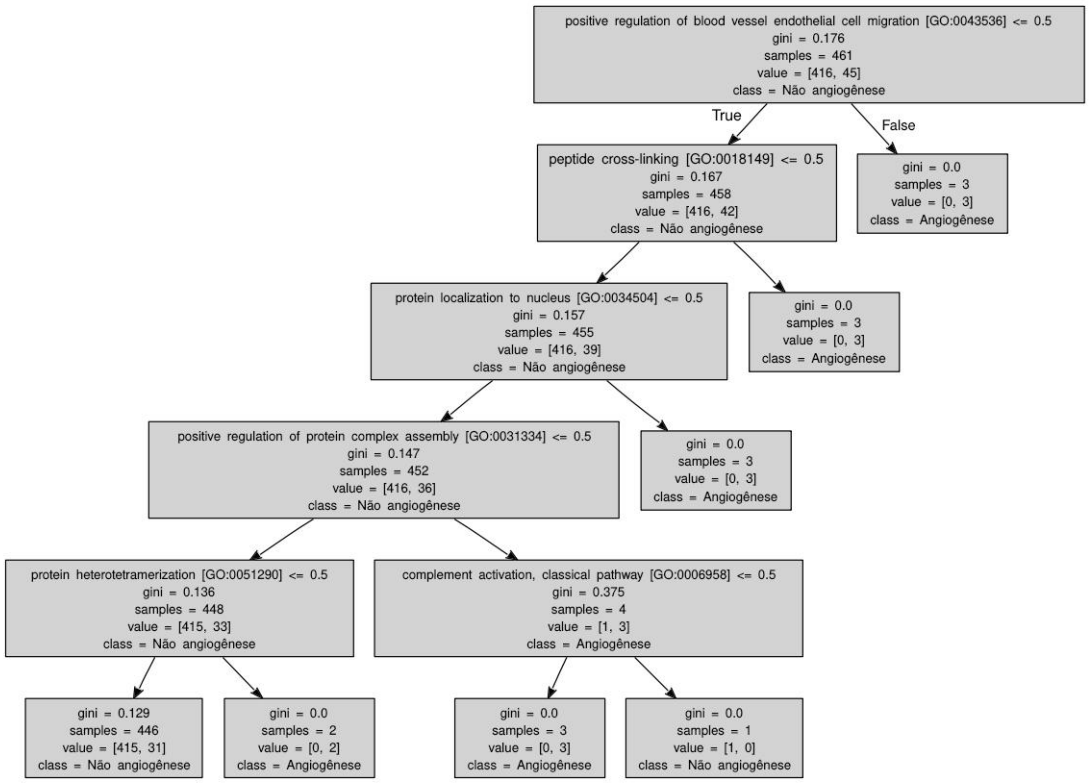

Figura 11 – Árvore mais ampla de funções relacionadas a angiogênese

Na árvore tem-se como mais importante uma função obviamente ligada à angiogênese, mas o seguinte, ligação cruzada de peptídeo, aparece no artigo (WANG; LI; CHEN, 2018). Este mostra uma aplicação de redes neurais analisando co-expressão para marcadores do melanoma e a ligação cruzada de peptídeo aparece como importante no desenvolvimento de melanoma metastático ao ser expresso com cornificação e queratinização induzida (GO:0070268) e regulação da concentração de íons de cálcio citosólico (GO:0051480).

Mas essas árvores não conseguiram atingir uma representatividade grande, os nós sempre dividem o grupo de amostrar em dois, três itens no máximo. Isso significa que por mais que apareçam como mais importantes na hierarquia da árvore, os nós superiores não tem muita informação adicionada em relação aos seguintes.

7.2.2 Árvore de dados clínicos

A última estratégia consistia em usar os dados clínicos dos pacientes para criar uma árvore que classificava sobre a sobrevivência destes, o objetivo foi tentar destacar quais

características são mais importantes para o nível de perigo de morte. A árvore resultante foi a mostrada na figura 12.

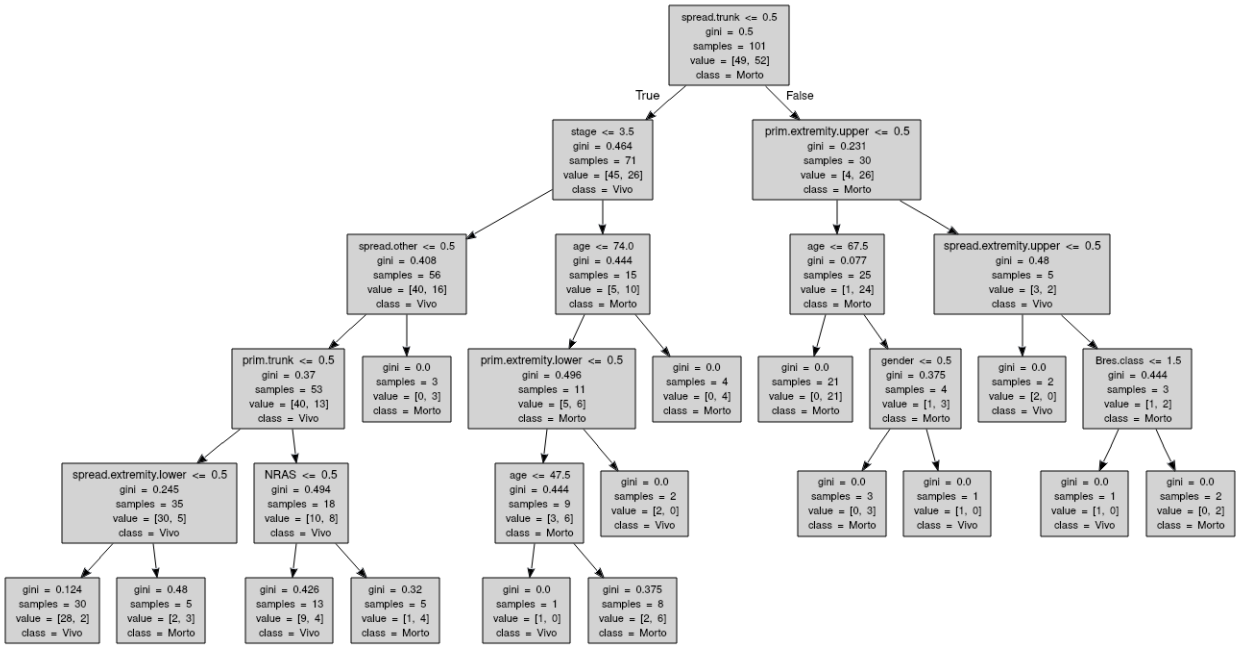

Figura 12 – Árvore de sobrevivência

Esta árvore consegue atingir um valor mais interessante de representatividade, já que as amostras são mais bem divididas. O dado mais importante para a decisão de sobrevivência aparece como sendo a difusão para o tronco, a idade mostra importância ao aparecer nos primeiros níveis. Interessante notar também que aparentemente as anotações dos oncogenes não demonstram tanto valor estatístico.

8 CONCLUSÃO

Tem muito campo para descoberta sobre o funcionamento de tumores, já que a doença se vale de muitas ferramentas complexas para seu desenvolvimento. Caminhos que façam vir à tona proteínas importantes no mar de proteínas que interagem no mecanismo celular podem trazer luz às pesquisas biológicas.

Inicialmente com o foco em estruturar as proteínas foram descobertas dimensões de pontos que trouxeram grande complexidade na retirada de informações. Mas com o objetivo de descobrir os pontos fora da curva, e ao usar métodos de redução de dimensionalidade, se possibilitou a marcação de algumas proteínas.

Ao tentar outras formas de interpretar, o problema foi montado mais diretamente, na pesquisa de quais funções biológicas são importantes para o desenvolvimento mortal do câncer. Com esse objetivo tentou-se usar das proteínas como sendo fonte de informação sobre tais funções, mas não se obteve muito sucesso em termos de representatividade.

Por outro lado, quando se pensou nos dados clínicos como sendo fonte primária dos treinamentos, houve uma representatividade maior. É interessante pensar na possibilidade de avaliar probabilidades de sobrevivência usando dados facilmente coletados dos pacientes, o que pode servir para tomada de decisão.

8.1 PRÓXIMOS TRABALHOS

Primeiramente pode se ter um treinamento de DCT com maiores volumes de dados. Os resultados foram representativos, mas o espaço amostral continha 111 pacientes, quanto mais pacientes for possível avaliar, mais próximos à realidade os valores vão ser.

O mais importante das avaliações que podem ser feitas é a de proteomas de pacientes não metastáticos. No trabalho foram verificados apenas pacientes que tiveram metástase nos seus tumores, o que restringe o campo de alcance do método. Se houver uma nova recuperação de dados proteômicos de pacientes doentes, mas que não se espalharam, nasce a possibilidade de uma comparação que pode levar à informações sobre a raiz da metástase.

Também há a possibilidade de criar novas formas de interpretar os pontos, talvez criar métodos geométricos que carregam o valor semânticos dos resultados. E também se pensar em avaliar as proteínas que desapareceram, não apenas as que estão nos dados. Muitas vezes os genes cancerígenos agem desativando a expressão de proteínas, o que nesse trabalho não é avaliado, mas seria bastante interessante verificar.

REFERÊNCIAS

BETANCOURT, L. H. et al. Improved survival prognostication of node-positive malignant melanoma patients utilizing shotgun proteomics guided by histopathological characterization and genomic data. **Scientific Reports**, Nature Publishing Group, v. 9, n. 1, dec 2019. ISSN 20452322. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6435712/>.

BRESLOW, A. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. **Annals of surgery**, Lippincott, Williams, and Wilkins, v. 172, n. 5, p. 902–908, 1970. ISSN 00034932.

BUCHAN, D. W. A.; JONES, D. T. Learning a functional grammar of protein domains using natural language word embedding techniques. **Proteins: Structure, Function, and Bioinformatics**, John Wiley and Sons Inc., v. 88, n. 4, p. 616–624, apr 2020. ISSN 0887-3585. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25842>.

CHO, S. Y. et al. Clinical Significance of the Thioredoxin System and Thioredoxin-Domain-Containing Protein Family in Hepatocellular Carcinoma. **Digestive Diseases and Sciences**, Springer New York LLC, v. 64, n. 1, p. 123–136, jan 2019. ISSN 15732568. Disponível em: <https://doi.org/10.1007/s10620-018-5307-x>.

DIMATOS, D. **MELANOMA CUTÂNEO NO BRASIL SKIN MELANOMA IN BRAZIL**. [S.l.], 2009. v. 38. Disponível em: <http://www.acm.org.br/revista/pdf/artigos/637.pdf>.

E. LANDMAN G., B. F. e. S. R. G. Estadiamento do Melanoma pela AJCC – 8ª edição - 2017. **MELANOMA**, n. 76, p. 3–7, 2017.

HAASS, N. K.; SMALLEY, K. S.; HERLYN, M. **The role of altered cell-cell communication in melanoma progression**. J Mol Histol, 2004. 309–318 p. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/15339050/>.

HANAHAN, D.; WEINBERG, R. A. The hallmarks of cancer. Elsevier, v. 100, n. 1, p. 57–70, jan 2000. ISSN 00928674. Disponível em: <http://www.cell.com/article/S0092867400816839/fulltexthttp://www.cell.com/article/S0092867400816839/abstracthttps://www.cell.com/cell/abstract/S0092-8674(00)81683-9>.

HE, G. et al. Cysteine-Rich Intestinal Protein 1 Silencing Inhibits Migration and Invasion in Human Colorectal Cancer. **Cellular Physiology and Biochemistry**, S. Karger AG, v. 44, n. 3, p. 897–906, nov 2017. ISSN 14219778. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29179181/>.

HIE, B. et al. Learning the language of viral evolution and escape. **Science**, American Association for the Advancement of Science (AAAS), v. 371, n. 6526, p. 284–288, jan 2021. ISSN 0036-8075. Disponível em: <http://science.sciencemag.org/>.

HOSAKA, Y. et al. Inhibition of invasion and experimental metastasis of murine melanoma cells by human soluble thrombomodulin. **Cancer Letters**, Cancer

Lett, v. 161, n. 2, p. 231–240, dec 2000. ISSN 03043835. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/11090974/>.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. International Conference on Learning Representations, ICLR, dec 2014.

MATTIUZZI, C.; LIPPI, G. Current cancer epidemiology. **Journal of Epidemiology and Global Health**, Atlantis Press International, v. 9, n. 4, p. 217–222, dec 2019. ISSN 22106014. Disponível em: <https://www.atlantis-press.com/journals/jegh/125919425>.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. International Conference on Learning Representations, ICLR, jan 2013.

MOCCIA, F. et al. **Endothelial Ca2+ signaling, angiogenesis and vasculogenesis: Just what it takes to make a blood vessel**. MDPI AG, 2019. Disponível em: </pmc/articles/PMC6721072/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6721072/>.

MORTON, D. L. et al. Multivariate analysis of the relationship between survival and the microstage of primary melanoma by clark level and breslow thickness. **Cancer**, John Wiley & Sons, Ltd, v. 71, n. 11, p. 3737–3743, jun 1993. ISSN 1097-0142.

MUNARON, L. Multilevel complexity of calcium signaling: Modeling angiogenesis. **World Journal of Biological Chemistry**, Baishideng Publishing Group Inc., v. 3, n. 6, p. 121, 2012. ISSN 1949-8454. Disponível em: </pmc/articles/PMC3421110/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3421110/>.

NG, A. **CS294A Lecture notes Sparse autoencoder**. [S.l.], 2011. Disponível em: <https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf>.

NUNEZ, L. **Malignant Melanoma**. 2019. Disponível em: <http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD009630>.

PRIYAM, A. et al. Comparative Analysis of Decision Tree Classification Algorithms. v. 3, n. 2, 2013. Disponível em: <http://inpressco.com/category/ijcet>.

SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. **IEEE Transactions on Systems, Man and Cybernetics**, v. 21, n. 3, p. 660–674, 1991. ISSN 21682909.

TRIOLA, M. **Introdução à estatística**. [S.l.]: LTC, 2005. ISBN 9788521614319.

WANG, L.-x.; LI, Y.; CHEN, G.-z. Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. **PLOS ONE**, Public Library of Science, v. 13, n. 1, p. e0190447, jan 2018. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0190447>.

WU, C. C. et al. p38γ regulates UV-induced checkpoint signaling and repair of UV-induced DNA damage. **Protein and Cell**, Higher Education Press, v. 1, n. 6, p. 573–583, 2010. ISSN 16748018. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3002048/>.

ZHANG, L. Z. et al. CRIP1 promotes cell migration, invasion and epithelial-mesenchymal transition of cervical cancer by activating the Wnt/β-catenin signaling pathway. **Life Sciences**, Elsevier Inc., v. 207, p. 420–427, aug 2018. ISSN 18790631. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29959029/>.

GLOSSÁRIO

angiogênese formação de novos vasos sanguíneos.

API *Application Programming Interface*, padrões que possibilitam interação entre códigos diferentes.

apoptose processo de morte celular programada.

classificador procedimento que decide quais elementos de uma população pertence a uma determinada classe.

enterócito tipo de célula epitelial da camada superficial dos intestinos.

fosfatase enzimas que removem um grupo fosfato do seu substrato.

fosfolipídio classe de lipídeos que são um dos principais componentes da membrana plasmática da célula.

genótipo composição genética do indivíduo.

hipoxia ausência de oxigênio suficiente nos tecidos para manter as funções corporais.

histopatológico estudo de como uma doença específica afeta um conjunto de células.

interferon proteína produzida pelos leucócitos e fibroblastos para interferir na replicação de invasores ou tumores.

melanócito células produtoras de melanina.

metástase evasão do cancêr de seu tecido original e proliferação em outros tipos de tecido.

neoplasia proliferação desordenada de células no organismo.

oncogene genes relacionados com o aparecimento e crescimento de tumores.

ontologia conjunto de conceitos dentro de um domínio e os relacionamentos entre estes.

peroxidase grupo de enzimas oxirredutases que oxidam substratos orgânicos.

prognóstico conhecimento ou juízo antecipado, prévio, feito pelo médicos.

protease enzimas que quebram ligações peptídicas entre os aminoácidos das proteínas.

proteoma conjunto de proteínas e variantes de proteínas que podem ser encontrados numa célula específica quando esta está sujeita a um certo estímulos.

transcriptoma Coleção de RNAs (transcritos) presentes em uma célula/tecido num dado momento.

APÊNDICES

APÊNDICE A – CÂNCER

O câncer se caracteriza como um crescimento desordenado de células, que pode eventualmente se espalhar por vários tecidos diferentes, causando até a morte do paciente. Mas para que uma célula cancerígena se multiplique a ponto de causar problemas ao organismo ela precisa de várias modificações no seu funcionamento que possibilitem sua sobrevivência aos mecanismos de controle do corpo e a capacidade de usar processos externos para sua estabilização.

Um artigo que marcou a pesquisa desses processos que possibilitam o desenvolvimento cancerígeno é o artigo (HANAHAN; WEINBERG, 2000) que fala sobre as grandes marcas descobertas até o início do ano 2000. Há mais de cem tipos distintos de câncer e subtipos de tumores que podem ser encontrados em variados órgãos, o que traz muitas perguntas sobre formas para cada tipo de célula se multiplicar em cada ambiente, interagindo com ele.

Hanahan (HANAHAN; WEINBERG, 2000) sugere que dentre o vasto catálogo de genótipos de células cancerígenas existem seis alterações essenciais na sua fisiologia que ditam o crescimento maligno (figura 13): auto-suficiência em sinalização de crescimento, insensibilidade a sinalizações inibitórias de crescimento, evasão da apoptose, potencial replicativo ilimitado, angiogênese sustentada e invasão de tecido para a metástase.

A.1 AUTO-SUFICIÊNCIA EM SINALIZAÇÃO DE CRESCIMENTO

Para que células normais mudem de um estado de equilíbrio para a proliferação são requeridos sinais mitógenos de crescimento. Mitógenos são substâncias transmitidas por receptores transmembrana que desencadeiam mais moléculas de sinalização: de fatores de crescimento difuso, de componentes de matriz extracelular e de interação entre células. Até onde se sabe não há como uma célula normal se multiplicar sem esses sinais.

Muitos oncogenes agem mimetizando esses sinais, assim possibilitando a proliferação tumoral, que mostra uma incomum independência desses sinais mitógenos. A conclusão é que essas células doentes geram sua própria sinalização de crescimento, assim conseguindo essa redução de dependência. Essa liberdade adquirida de sinais exógenos irrompe um mecanismo homeostático criticamente importante que normalmente garante o equilíbrio de células num tecido.

A autonomia a esses sinais foi a primeira descoberta claramente definida pelos pesquisadores, principalmente pela prevalência de oncogenes dominantes. Há três alterações estratégicas que ficaram evidentes: de sinais de crescimento, na transdução transcelular desses sinais e nos circuitos que traduzem esses sinais para ação.

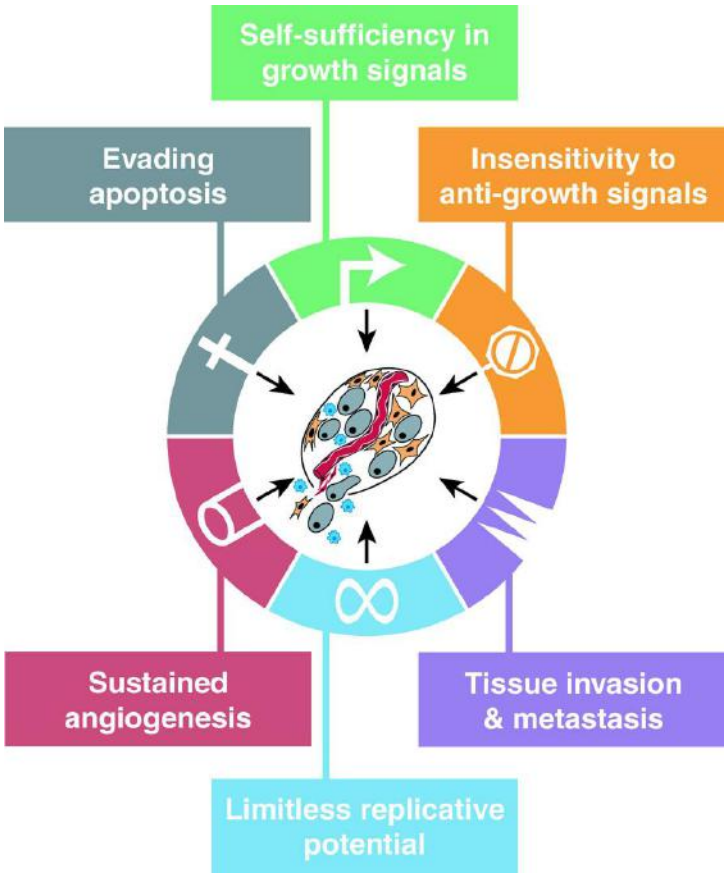

Figura 13 – Capacidades adquiridas do câncer (HANAHAN; WEINBERG, 2000)

Enquanto a maioria dos fatores de crescimento mitógenos (GFs) são produzidos em uma célula para afetar outra, muitas células cancerígenas adquirem a habilidade de sintetizar GFs para os quais elas são responsivas, criando um ciclo de sinalização de retroalimentação positiva, muitas vezes conhecido por estimulação autócrina. Dois exemplos ilustrativos são o PDGF (fator de crescimento derivado de plaquetas) e o TGFα (fator de crescimento transformador alfa) por glioblastomas e sarcomas, respectivamente (HANAHAN; WEINBERG, 2000).

Outros alvos da desregulação causada pela patogênese do tumor são os receptores de superfície que transduzem sinais estimulatórios de crescimento. Receptores de GF muitas vezes são superexpressos, o que faz a célula se tornar muito responsiva aos níveis ambiente desses fatores de crescimento, níveis que normalmente não desencadeariam a multiplicação celular. Por exemplo o EGF-R/erbB (receptor do fator de crescimento epidérmico) é regulado positivamente nos tumores de estômago, cérebro e mama, enquanto o HER2/neu (receptor do fator de crescimento da epiderme humana 2) tem expressão aumentada nos tumores de estômago e mama.

As células cancerígenas também mudam os receptores extracelulares (integrinas), favorecendo os de crescimento. Ambos os receptores GF ativados por ligação e integrinas de crescimento engajados com componentes da matriz extracelular podem ativar a via

metabólica SOS-Ras-Raf-MAP kinase que promove a mitose.

Os mecanismos mais complexos de sinalização de crescimento derivam de alterações nos componentes do circuito citoplasmático jusante que recebe e processa esses sinais emitidos pelos receptores GF e integrinas. Em cerca de 25% dos tumores humanos, as proteínas Ras estão em formas estruturalmente alteradas a ponto de liberarem um fluxo de sinais mitógenos sem estimulação dos seus reguladores montantes (HANAHAN; WEINBERG, 2000).

Por mais que a aquisição de autonomia à sinalização de crescimento possa ser conceitualmente satisfatória ela também é muito simplista. A desregulação de crescimento pode também ser explicada ao entender as contribuições das células auxiliares presentes no tumor, os aparentemente normais expectadores, como os fibroblastos e células endoteliais que fazem um papel importante na proliferação das células tumorais. Por exemplo, suspeita-se (HANAHAN; WEINBERG, 2000) que os sinais de crescimento que geram a multiplicação se originam nas células estromais componentes da massa do tumor.

A.2 INSENSIBILIDADE A SINALIZAÇÕES INIBITÓRIAS DE CRESCIMENTO

Dentro de um tecido normal muitos sinais antiproliferativos operam para manter a homeostase celular. Esses sinais, como seus homólogos positivos, são recebidos pelos receptores da superfície celular transmembrana, acoplados em circuitos intracelulares de sinalização. Os sinais de anticrescimento bloqueiam a multiplicação celular com dois mecanismos:

- Forçar as células a saírem de um estado proliferativo para um de repouso (G0) do qual elas podem reemergir numa ocasião futura onde os sinais extracelulares permitirem.
- As células também podem ser induzidas a renunciar permanentemente do seu potencial reprodutivo ao ser colocada em estados pós mitose, usualmente associados à aquisição de feições associadas à diferenciação.

Muito do mecanismo que as células normais usam para responder a sinais anticrescimento está associado ao ciclo celular, especificamente os componentes que governam a passagem pela fase de restrição (G1). As células monitoram o ambiente externo nesse período, e baseado nesses sinais é decidido se vai entrar num estado proliferativo, quiescente ou pós mitótico. A nível molecular, muitos, talvez todos, os sinais antiproliferativos são afunilados através da proteína do retinoblastoma (pRb) e suas duas relacionadas p107 e p130. Quando hipofosforilada, a pRb bloqueia a multiplicação celular sequestrando e alterando a função dos fatores de transcrição E2F que controlam a expressão de bancos de genes essenciais para a progressão da fase G1 para a fase de síntese (S).

Disrupção da via metabólica da pRb libera os E2Fs e assim libera a proliferação celular, criando células insensíveis à fatores anticrescimento. Os efeitos da molécula de sinalização $TGF\beta$ são os mais bem documentados, um deles é prevenir a fosforilação que desativa a pRb, assim bloqueando o avanço à fase G1. Em alguns tipos de célula a $TGF\beta$ suprime a expressão do gene c-myc, que regula o mecanismo da fase G1. Mais diretamente, a $TGF\beta$ causa a síntese das proteínas p15 e p21, que bloqueiam as ciclinas (complexos CDK responsáveis pela fosforilação da pRb)

O circuito de sinalização da pRb, governado pela $TGF\beta$ e outros fatores externos, podem ser alterados de várias formas nos diferentes tipos de tumores:

- Alguns perdem a responsividade ao $TGF\beta$ através da regulação negativa de seus receptores, outros mostram receptores mutantes, disfuncionais.
- A proteína citoplasmática Smad4, que transduz sinais dos receptores $TGF\beta$ para alvos jusantes, pode ser eliminada através da mutação do seu gene codificante.
- O lócus de codificação da proteína p15 pode ser deletado.
- O alvo jusante da p15, CDK4, também pode ser afetado, não sendo mais responsivo às suas ações inibitórias. Isso acontece com mutações que criam substituições dos aminoácidos no domínio de interação INK4B, os complexos de ciclinas D:CDK4 resultantes são liberados para desativar a pRb por hiperfosforilação (HANAHAN; WEINBERG, 2000).
- A pRb funcional, alvo fim dessa via metabólica, pode ser perdida através da mutação do seu gene. Também em alguns tumores induzidos por vírus, notavelmente o carcinoma cervical, a função da pRb é eliminada através de sequestro por oncoproteínas, como a E7, do papilomavírus humano.
- As células cancerígenas também desligam a expressão de integrinas e outras moléculas de adesão celular que enviam os sinais anticrescimento, favorecendo no seu lugar a transmissão de sinais favoráveis ao crescimento.

A proliferação celular depende de mais do que da fuga dos sinais citostáticos anticrescimento. Os tecidos humanos também constrangem a multiplicação celular ao instruir as células a entrarem em estado pós mitótico, estados diferenciados, usando diversos mecanismos ainda não completamente compreendidos, mas é aparente que o tumor tem várias estratégias para evadir essa diferenciação terminal.

Uma estratégia para a fuga da diferenciação envolve diretamente o oncogene c-myc, que codifica um fator de transcrição. Durante o desenvolvimento normal, a ação estimulatória de crescimento do Myc, em associação com o fator Max, pode ser substituída por

complexos alternativos de Max com um grupo de fatores de transcrição Mad, os complexos Mad-Max extraem sinais indutores de diferenciação. Entretanto, a superexpressão da c-myc, como visto em vários tumores, pode reverter esse processo, desbalanceando a favor de complexos Myc-Max, assim impedindo diferenciação e promovendo multiplicação. Durante a carcinogênese do cólon humano, a inativação da via metabólica $APC/\beta-catenin$ bloqueia a passagem de enterócitos nas criptas colônicas para um estado diferenciado, pós mitótico (HANAHAN; WEINBERG, 2000).

A.3 EVASÃO DA APOPTOSE

Para que a população celular tumoral cresça em números é necessário mais do que um aumento na sua capacidade proliferativa, estudos em ratos e células de cultura, tal qual em análises descritivas de estágios da carcinogênese humana são evidências de que a resistência à apoptose (morte celular programada) é uma marca da maioria, se não todos, os tipos de câncer.

Observações de pesquisas indicam que a programação apoptótica é presente em forma latente em virtualmente todos os tipos de células do corpo (HANAHAN; WEINBERG, 2000). Uma vez ativadas por uma série de sinais fisiológicos, esse programa entra em curso numa série coreografada de passos. Membranas celulares são rompidas, os esqueletos citoplasmáticos e nucleares são quebrados, o citosol é extrusado, os cromossomos degradados e o núcleo é fragmentado, tudo em um intervalo de trinta a cento e vinte minutos. No fim, o corpo murcho da célula é absorvido pelas células próximas e desaparece, tipicamente no período de um dia (HANAHAN; WEINBERG, 2000).

O maquinário apoptótico pode ser dividido em duas classes de componentes: os sensores e os efetores. Os sensores são responsáveis por monitorar o ambiente extracelular para condições que influenciam se uma célula deve morrer ou continuar a viver. Esses sinais regulam a segunda classe de componentes, que ativamente destroem a célula. Exemplos desses pares ligantes-receptores são os sinais de sobrevivência entregues pelos IGF-1/IGF-2 através do seu receptor IGF-1R, e pelo IL-3 e seu receptor cognato IL-3R (HANAHAN; WEINBERG, 2000). Sinais de morte são transmitidos pelo ligante FAS, conectando ao receptor FAS e pelo $TNF\alpha$ conectando o TNF-R1 (HANAHAN; WEINBERG, 2000). Sensores intracelulares também monitoram a saúde da célula e ativam o processo de apoptose se detectam anormalidades, como dano ao DNA, desequilíbrio de sinalização causados por ações de oncogenes, insuficiência de fatores de sobrevivência, como hipoxia.

Muitos dos sinais que induzem apoptose convergem na mitocôndria, que responde liberando citocromo C, um potente catalisador da apoptose (HANAHAN; WEINBERG, 2000). Membros da família de proteínas Bcl-2 tem ambos os componentes proapoptóticos (Bax, Bak, Bid, Bim) e antiapoptóticos (Bcl-2, Bcl-XL, Bcl-W), agem em parte por

governar a sinalização de morte mitocondrial através da liberação do citocromo C. A proteína p53 suprime tumores induzindo apoptose ao regular positivamente a expressão da Bax ao detectar dano ao DNA, a Bax, por sua vez estimula a mitocôndria a liberar o citocromo C.

Os principais efetores da apoptose incluem um grupo de proteases conhecidas por caspases (HANAHAN; WEINBERG, 2000). Duas caspases de entrada, -8 e -9, são ativadas pelos receptores de morte FAS e citocromo C, respectivamente. Essas caspases proximais engatilham a ativação de dúzias de caspases que executam o programa de morte, através da destruição seletiva de estruturas e organelas subcelulares, inclusive o genoma.

A possibilidade de que a apoptose fosse uma barreira para o câncer foi levantada primeiramente em 1972, quando Kerr, Wyllie e Currie descreveram uma morte massiva numa população de células que cresciam rapidamente, e tumores dependentes de hormônio gerando de uma retirada desses hormônios (HANAHAN; WEINBERG, 2000). A descoberta do oncogene bcl-2 pela sua regulação positiva via translocação cromossomal no linfoma folicular e seu reconhecimento como antiapoptótico (HANAHAN; WEINBERG, 2000) começou a investigação da apoptose no câncer a nível molecular. Quando o bcl-2 é expresso juntamente ao oncogene myc, ele é capaz de promover formação de linfomas de célula B, aumentando a sobrevivência dos linfócitos, não pela proliferação induzida pelo myc (HANAHAN; WEINBERG, 2000).

Mais informações sobre a interação myc-bcl-2 foram descobertas depois ao estudar os efeitos do oncogene myc numa cultura de fibroblastos com soro reduzido. A apoptose foi induzida amplamente nas células com myc expresso em células de soro reduzido, essas mortes podem ser revogadas por fatores externos de sobrevivência (e.g. IGF-1), por superexpressão forçada de Bcl-2 ou a proteína relacionada Bcl-X ou pelo rompimento do ciclo de sinalização de morte FAS (HANAHAN; WEINBERG, 2000). Coletivamente, os dados indicam que o programa de morte celular pode ser ativado por um oncogene superexpresso.

Resistência à apoptose pode ser adquirida pelas células através de uma série de estratégias. Certamente a mais comum das formas de se perder a regulação apoptótica envolve mutação do gene supressor de câncer, p53. A inativação do seu produto, a proteína p53, é vista em mais de 50% dos cânceres humanos e resulta na remoção de um componente chave que é sensor de danos ao DNA e ativador da cascata de efeitos apoptóticos (HANAHAN; WEINBERG, 2000). Adicionalmente, a via metabólica PI3 kinase–AKT/PKB, que transmite sinais de sobrevivência antiapoptóticos é provavelmente envolvida numa fração substancial dos tumores humanos. Esse circuito de sinalização de sobrevivência é ativado por fatores extracelulares como o IGF-1/2 e o IL-3, por sinais intracelulares emanando do Ras ou pela perda do supressor de tumores pTEN, uma fosfatase fosfolipídio que normalmente atenua o sinal de sobrevivência AKT (HANAHAN; WEINBERG, 2000). Um mecanismo para revogar o sinal de morte FAS foi revelado de uma alta fração

de células de carcinomas de pulmão e cólon: um chamariz não sinalizante receptor de FAS ligante é regulado positivamente, atraindo o sinal indutor de morte para longe do seu receptor FAS (HANAHAN; WEINBERG, 2000). É virtualmente esperado que todas as células cancerígenas tenham mecanismos para possibilitar a evasão da apoptose.

A.4 POTENCIAL REPLICATIVO ILIMITADO

Inicialmente se acreditava que a proliferação desregulada era o suficiente para possibilitar a geração de grandes populações de tumores macroscópicos, mas pesquisas mais recentes indicam que apenas a perturbação na sinalização intercelular não garante o crescimento expansivo do tumor (HANAHAN; WEINBERG, 2000). Muitos, talvez todos, os tipos de células mamíferas tem programas que limitam a sua multiplicação, esses programas operam de forma diferente das sinalizações célula a célula, como as vias metabólicas descritas acima.

O trabalho de Hayflick demonstra que células em cultura tem um potencial replicativo finito (HANAHAN; WEINBERG, 2000), uma vez que as populações atingem um certo número, elas param de se multiplicar (um processo conhecido por senescência). Esse estado crítico é caracterizado por morte massiva da célula, neste momento que ocasionalmente emerge a célula variante (1 em 10^7) que adquire a habilidade de multiplicar sem limite (traço conhecido por imortalização) (HANAHAN; WEINBERG, 2000).

Provocativamente a maioria dos tipos de células tumorais aparentam estar imortalizadas, sugerindo que o potencial replicativo ilimitado é um fenótipo adquirido *in vivo* durante a progressão do tumor e foi essencial para o estado de crescimento maligno do tumor (HANAHAN; WEINBERG, 2000). Isso sugere que em algum ponto durante o curso da progressão do tumor, as populações pré malignas evoluindo esgotaram o legado de todas as células que não são capazes de atravessar a barreira da mortalidade.

Telômeros são estruturas da extremidade dos cromossomos que são encurtadas a cada divisão celular, ao atingir um tamanho limite a replicação é impossibilitada. A manutenção do telômero é virtualmente evidente em todos os tipos de células malignas (HANAHAN; WEINBERG, 2000). 85% a 90% obtém sucesso em fazê-lo ao regular positivamente a expressão da enzima telomerase que adiciona repetições de hexanucleotídeos no final dos telômeros, enquanto o restante inventou uma forma de ativar um mecanismo (conhecido por ALT) que parece manter os telômeros através da recombinação de trocas intercromossomais (HANAHAN; WEINBERG, 2000). Ambos os mecanismos parecem ser fortemente reprimidos na maioria das células humanas, a fim de bloquear seu potencial de replicação infinito.

Pistas adicionais da importância da manutenção do telômero vêm da análise de ratos com falta da função da telomerase. Por exemplo, ratos carregando o nocaute homozigoto do inibidor do ciclo celular p16 são suscetíveis a tumores, particularmente quando ex-

postos à carcinogenes, os tumores que surgem mostram elevada atividade da telomerase, comparativamente. Quando carcinogenes são aplicados a ratos sem p16 que também faltavam telomerase, a incidência de tumores era reduzida, concomitante com a substancial baixa de telômeros e morte nos tumores que apareciam (HANAHAN; WEINBERG, 2000).

A manutenção dos telômeros é um componente chave para a capacidade de replicação ilimitada, mas ainda não se sabe exatamente sobre a outra questão, a fuga da senescência celular, que pode representar um passo essencial na evolução do tumor, dado que é uma barreira crítica. Mas também se considera um modelo alternativo igualmente plausível: a senescência pode ser um evento da cultura celular e não reflete um fenótipo de células vivendo em tecidos, e assim não seria um problema para o tumor *in vivo*. A resolução desse dilema pode ser importante no entendimento do potencial replicativo de cânceres.

A.5 ANGIOGÊNESE SUSTENTADA

Oxigênio e nutrientes supridos por uma vasculatura são cruciais para a sobrevivência e função celular, virtualmente obrigando células num tecido a residirem $100\mu m$ de um vaso sanguíneo capilar. Dada essa dependência de capilares próximos, parece plausível que células se proliferando teriam uma habilidade intrínseca de induzir a criação de novos vasos sanguíneos. Mas as evidências mostram o oposto, células de proliferação descontrolada de lesões inicialmente não tem a habilidade angiogênica, limitando sua capacidade de expansão. Para se atingir progressão, neoplasias incipientes devem desenvolver tal habilidade (HANAHAN; WEINBERG, 2000).

O balanço de sinais positivos e negativos incentivam ou bloqueiam a angiogênese. Os sinais iniciadores são exemplificados pelo fator de crescimento vascular endotelial (VEGF) e fatores de crescimento de fibroblastos (FGF1/2). Cada um se conecta a receptores de tirosina quinase dispostos nas células endoteliais. O inibidor inicial de angiogênese é a trombospondina 1, que se conecta ao CD36, um receptor transmembrana das células endoteliais acoplado às tirosina quinases. Há mais de duas dúzias de fatores indutores de angiogênese conhecidos, e número similar de proteínas inibidoras endógenas.

Sinalização de integrinas também contribui para o balanço regulatório. Vasos em equilíbrio expressam uma classe de integrinas, enquanto os que estão brotando expressam outras classes. Proteases extracelulares são física e funcionalmente conectadas às integrinas pró angiogênese e ambas ajudam a ditar a capacidade invasiva das células angiogênicas endoteliais.

Evidência experimental para a importância de induzir e sustentar a angiogênese é extensiva e convincente (HANAHAN; WEINBERG, 2000). Prova do princípio molecular veio quando anticorpos anti-VEGF se provaram capazes de parar a vascularização e crescimento de tumores subcutâneos em ratos, ou com a versão interferente dominante do VEGF, receptor 2 (flk-1) (HANAHAN; WEINBERG, 2000), resultados que motivaram o

desenvolvimento de inibidores VEGF/VEGF-R.

A angiogênese pode ser encontrada em lesões pré malignas de cânceres humanos do cérvix, mama e pele (melanoma), é esperado que sua indução vai se provar acontecer nos estágios iniciais a médios da doença. Tumores parecem ativar a angiogênese ao mudar o equilíbrio dos seus indutores em relação aos seus inibidores, na qual tem-se de estratégia comum alterar a transcrição genética. Muitos tumores mostram aumento na expressão dos VEGF ou FGFs comparados às células normais. Em outros, a expressão de inibidores endógenos como a trombospondina 1, ou o interferon β sendo negativamente regulado. Além disso, pode-se ocorrer ambas transições de forma ligada (HANAHAN; WEINBERG, 2000).

Os mecanismos de desbalanceamento dos reguladores de angiogênese ainda não são completamente entendidos, mas num exemplo bem documentado a trombospondina 1 foi encontrada positivamente regulada pela proteína p53 em alguns tipos de célula. Consequentemente, a perda da função da p53, o que ocorre na maioria dos tumores humanos, pode causar a queda dos níveis de trombospondina 1, liberando as células endoteliais de estados de inibição (HANAHAN; WEINBERG, 2000). O gene VEGF também fica sob controle de transcrição complexo. Por exemplo, a ativação do oncogene ras ou a perda do gene supressor de tumores VHL em alguns tipos de célula regulam positivamente a expressão do VEGF.

Outra dimensão da regulação emerge na forma de proteases, que podem controlar a biodisponibilidade dos ativadores e inibidores angiogênicos. Uma variedade de proteases podem liberar os FGFs básicos estocados na matriz extracelular, onde a plasmina (componente pró angiogênico do sistema de coagulação) pode se fragmentar num inibidor chamado angiostatina (HANAHAN; WEINBERG, 2000). A expressão coordenada de moléculas sinalizadoras e sua modulação por proteólise, parece refletir na regulação homeostática da angiogênese em tecidos normais.

A.6 INVASÃO DE TECIDO E METÁSTASE

Eventualmente, no desenvolvimento da maioria dos tipos de cânceres humanos, massas primárias de tumor produzem células pioneiras que se movem para fora, invadem tecidos adjacentes e mais, atingem também locais distantes onde podem suceder em criar novas colônias. Esses estabelecimentos distantes (metástase) são causa de 90% das mortes humanas para o câncer (HANAHAN; WEINBERG, 2000). Como na formação da massa primária, sucesso na metástase depende de adquirir todas as cinco capacidades anteriores. Mas quais mudanças celulares possibilitam a aquisição dessa capacidade final?

Várias classes das proteínas envolvidas na amarração de células ao seu entorno num tecido são alteradas em células que possuem capacidades invasivas. As proteínas afetadas incluem moléculas de adesão célula a célula (CAMs), que tem de membros notáveis as

imunoglobulinas e famílias de caderinas dependentes de cálcio, e integrinas, que conecta as células aos substratos da matriz extracelular.

A alteração cancerígena de interação célula com ambiente mais amplamente observada envolve a E-caderina, molécula de interação intercelular homotípica costumeiramente expressa em células epiteliais. Acoplamento de células adjacentes por pontes de E-caderina resulta na transmissão de sinais, inclusive os anti crescimento, através de contato citoplasmático com a β-catenina para circuitos intracelulares que incluem o fator de transcrição Lef/Tcf (HANAHAN; WEINBERG, 2000). A função da E-caderina parece ser perdida na maioria dos cânceres epiteliais, por mecanismos que incluem inativação por mutação de ambos os genes da E-caderina e da β-catenina, repressão transcricional ou proteólise do domínio extracelular da caderina. Expressão forçada da E-caderina em células de câncer em cultura e num modelo transgênico de rato debilita os fenótipos metastáticos (HANAHAN; WEINBERG, 2000), assim a E-caderina serve como um supressor de ação ampla de invasão em cânceres epiteliais, e sua eliminação funcional representa um passo importante para essa capacidade.

Mudanças nas expressões das CAMs de famílias de imunoglobulinas também apresentam importante papel no processo de invasão e metástase. O caso mais claro envolve a N-CAM, que passa por mudança de expressão de uma isoforma altamente adesiva para uma forma pouco adesiva (eventualmente repulsiva) do tumor de Wilms, neuroblastoma e câncer de pulmão de células pequenas; tanto como sua redução de expressão geral nos cânceres invasivos de pâncreas e colorretais (HANAHAN; WEINBERG, 2000). Experimentos em ratos transgênicos apoiam o papel funcional da forma adesiva da N-CAM em suprimir metástase (HANAHAN; WEINBERG, 2000).

Mudanças na expressão das integrinas também são evidentes nas células invasivas e metastáticas. Sucesso na colonização de novos ambientes (locais e distantes) demanda adaptação, que é alcançada através de mudanças nos espectros de integrinas α ou β mostrado nas células migratórias. Essas modificações resultam subtipos diferentes de integrinas (do qual tem mais de 22) de diferentes substratos preferíveis. Assim, células cancerígenas facilitam a invasão ao mudar a expressão de suas integrinas daquelas que favorecem a matriz extracelular original para outras integrinas (e.g. $\alpha 3\beta 1$ ou $\alpha V\beta 3$) que preferencialmente se conectam aos componentes estromais degradados produzidos pelas proteases extracelulares (HANAHAN; WEINBERG, 2000). A expressão forçada de subunidades de integrinas pode induzir ou inibir comportamento metastático, o que é consistente com a importância desses receptores na determinação desse processo.

Tentativas de explicar os efeitos biológicos das integrinas em termos de um número menor de regras mecânicas foram frustradas pela enorme quantidade de diferentes genes das integrinas, e pelo número ainda maior de receptores heterodiméricos que resultam numa expressão combinatória de várias subunidades α e β, e ainda pela crescente evidência de sinais complexos emitidos de domínios citoplasmáticos dos receptores. Ainda sim existe

pouca dúvida de que esses receptores representam um papel central na capacidade de invasão e metástase (HANAHAN; WEINBERG, 2000).

O segundo parâmetro geral de capacidade metastática envolve proteases extracelulares. Genes de proteases são positivamente regulados, inibidores negativamente regulados e formas inativas de zimogênio de proteases são convertidas em enzimas ativas. Proteases que degradam matrizes são caracteristicamente associadas à superfície celular, acredita-se que podem facilitar a invasão de células cancerígenas no estromas próximos, através de paredes de vasos sanguíneos e através de camadas de células epiteliais normais. Apesar dessa noção, é difícil apontar a função de proteases separadamente por essa capacidade, dados seus importantes papéis em outras funções marcas importantes como a angiogênese e sinalização de crescimento (HANAHAN; WEINBERG, 2000).

Uma dimensão da complexidade é adicionada pelos múltiplos tipos de célula envolvidos na expressão e aparição da protease. Em muitos tipos de carcinomas, proteases que degradam matriz são produzidas não pelo câncer epitelial mas pela recrutação de células estromais e inflamatórias (HANAHAN; WEINBERG, 2000). Por exemplo, certos tipos de câncer induzem expressão da uroquinase (uPA) em células estromais em cultura, que conecta ao receptor de uroquinase (uPAR) das células cancerígenas.

APÊNDICE B – APRENDIZADO DE MÁQUINA.

A primeira ferramenta usada na tarefa de retirar informações dos dados, nesse trabalho, são as redes neurais. Inicialmente o trabalho consiste em resumir, enriquecer e visualizar os dados clínicos e histopatológicos colhidos dos pacientes com melanoma. Para isso foi usada uma implementação de *word2vec*, mas antes de passar por esse método serão apresentados conceitos que servirão de base para o seu entendimento.

B.1 REDES NEURAIS

A fim de se demonstrar uma rede neural será usado o exemplo mais simples possível, que é constituído de apenas um neurônio:

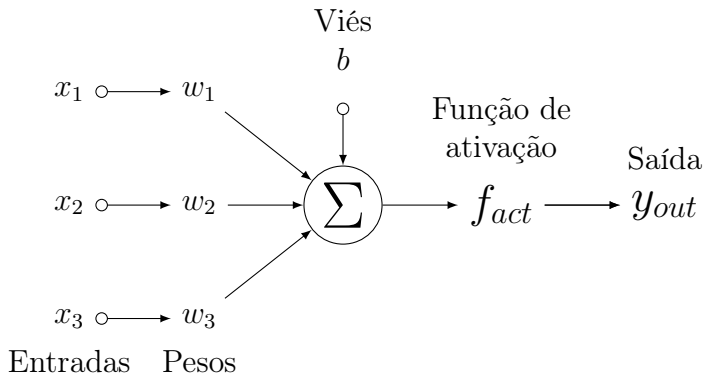

Figura 14 – Neurônio simples

Esse neurônio é uma unidade de computação que recebe como entrada os itens x_1, x_2, x_3. Estes são a descrição numérica do problema, cada um representa uma dimensão na qual os dados estão dispostos. Como um exemplo tem-se a cor de cada ponto de uma imagem.

Esses valores serão multiplicados pelos pesos w_1, w_2, w_3, que são as principais variáveis da rede, estes são ajustados durante o processo de aprendizagem.

Além das entradas também pode ser usado um viés b, que será somado aos valores anteriores, adicionando mais uma capacidade de adaptação no neurônio.

A soma desses valores será aplicada numa função de ativação, que determina se o neurônio vai disparar ou não, dependendo do que recebeu de entrada.

Um exemplo comum de função usada como ativação é a sigmoide:

$$f(z) = \frac{1}{1 + exp(-z)}$$

Que é menos sensível a pequenas mudanças que a primeira função utilizada para neurônios, a função degrau unitário:

$$f(z) = \begin{cases} 0 & \text{if } z <= 0 \\ 1 & \text{if } z > 0 \end{cases}$$

Usando este neurônio pode ser construída uma rede neural como sendo uma série de conexões destes. O padrão de conectividades e nós é conhecido como arquitetura. Sua arquitetura tem impacto na forma que a rede vai ser treinada e que tipo de informações podem ser retiradas dos nós.

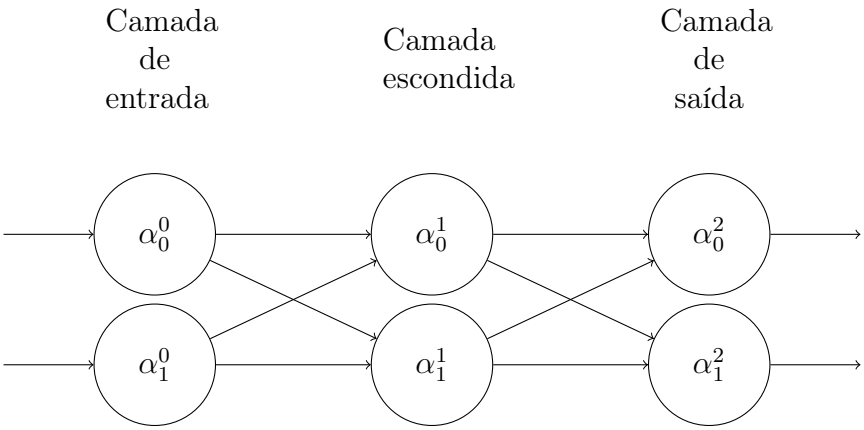

Figura 15 – Rede neural simples

Nesse exemplo temos uma rede com dois neurônios em cada camada. A entrada também pode ser representada com círculos. A camada escondida é onde tem-se os neurônios intermediários cujos valores não estão no conjunto de treino supervisionado, por isso "escondida". A camada de saída tem os neurônios que retornarão os valores reproduzidos pela rede.

B.2 AUTOENCODERS

Imagine agora um problema onde os dados de treinamento não tem rótulo. Um caso de representação ou classificação de imagens, áudio ou até mesmo texto.

Uma das formas de se lidar com esses dados é através da tentativa de replicação. Os autoencoders são uma arquitetura de redes neurais onde se usa o mesmo método de treino de uma rede supervisionada onde as saídas são iguais às entradas.

Em outras palavras, a rede tenta aproximar a função identidade (NG, 2011). Apesar de parecer vão, ao criar limites para a rede como limitar a quantidade de neurônios escondidos, pode-se descobrir informações interessantes sobre a rede.

Por exemplo, se houver na entrada uma imagem de dez pontos de largura e altura, serão necessários cem neurônios na camada de entrada. Para ser feita a replicação também haverá cem neurônios na saída. Mas se forem colocados dez neurônios na camada

escondida, o treinamento vai ser responsável por comprimir os dados da entrada em dez números, para ser capaz de reproduzí-los.

Se os dados forem completamente randômicos esse treino será muito difícil, mas se houver uma estrutura o algoritmo vai ser capaz descobrir essas relações.

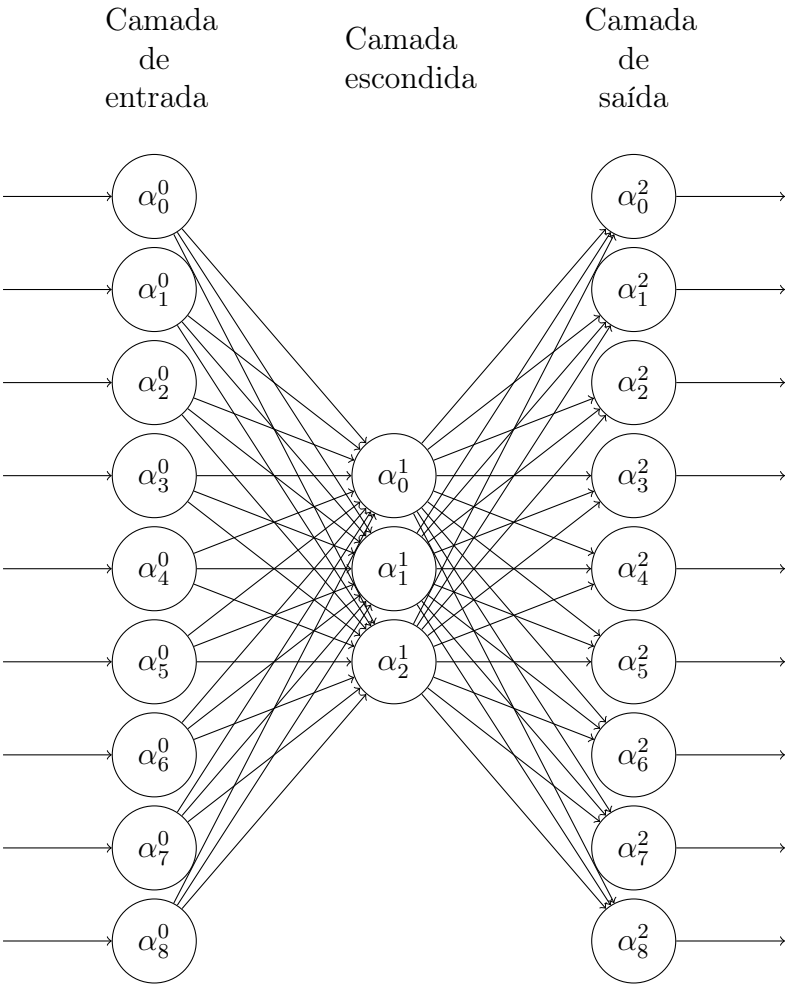

Figura 16 – Exemplo de autoencoder

As dimensões criadas pelas variáveis da camada escondida são chamadas de espaço latente. Existem formas de ordená-lo a fim de que se tenha uma reprodução previsível, possibilitando operações neste espaço que geram informação, como resultados novos ou interações entre itens.

Uma das arquiteturas que é orientada a se ter um espaço latente controlado é a de autoencoders variacionais (VAE). Estes autoencoders foram criados para executar inferências aproximadas eficientemente em modelos cujas variáveis latentes tem distribuição posterior intratável (KINGMA; WELLING, 2014). Ela pode ser usada para várias aplicações como reconhecimento, redução de ruído, visualização.

Processamento de linguagem natural é uma área onde o espaço latente tem um papel

muito importante dadas as várias representações da mesma ideia de formas diferentes. As representações e as relações podem ser demonstradas nesse espectro.

ANEXOS

ANEXO A – PROT2VEC

Código A.1 – Incorporação através de uma tabela de proteínas BC4r.xlsx - 1

```
import time
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import nltk as nltk
import re
import keras.backend as K
from keras.models import Sequential
from keras.layers import Dense, Embedding, Lambda, Flatten

from nltk.corpus import stopwords
nltk.download('stopwords')

BC4 = pd.read_excel('BC4r.xlsx')


def clean_tokens(tokens):
  sw = set(stopwords.words('english'))
  clean = []
  for token in tokens:
    lower = token.lower()
    if lower not in sw and lower:
      clean.append(lower.upper())
  return clean
```

Código A.2 – Incorporação através de uma tabela de proteínas BC4r.xlsx - 2

```
def gen_tokens_of_line(token_id, id_token, window_size, entry,
    prot_names, subcellular_loc, interacts_with, function_cc, polymorph,
    go_bio, go_cel, go_molec, domain_cc):
  words = []

  name = entry #Name
  words.append(name)

  desc = re.sub('\(|\)', '', re.sub(' [0-9].*?( |$)', '\g<1>',
     prot_names)) #Descriptions and names
  words = words + desc.split(' ')

  location = re.sub('\{|\}|\.', '', subcellular_loc[22:]) #Location
  words = words + location.split(' ')

  interaction = re.sub('Itself', name, re.sub(';', '', interacts_with))
     #Protein interaction
  words = words + interaction.split(' ')

  function = re.sub('\,|\{|\}|\.|\(|\)', '', function_cc[10:]) #Function
  words = words +function.split(' ')

  polymorphism = re.sub('\/', ' ', re.sub('\,|\{|\}|\.|\(|\)', '',
     polymorph[14:])) #Polymorphism
  words = words + polymorphism.split(' ')

  gobio = re.sub('\[|\]|;', '', go_bio) #GO Bio Process
  words = words + gobio.split(' ')

  gocel = re.sub('\[|\]|;', '', go_cel) #GO Cel Component
  words = words + gocel.split(' ')

  gomolec = re.sub('\[|\]|;', '', go_molec) #GO Molec Function
  words = words + gomolec.split(' ')

  domain = re.sub('DOMAIN: |\.|;|\{|\}|\(|\)', '', domain_cc) #Domain
  words = words + domain.split(' ')

  words = clean_tokens(words)

  for word in words:
    if(token_id.get(word) is None):
      token_id[word] = len(token_id)
      id_token[len(token_id)] = word

  window_size.append(len(words))
  return words
```

Código A.3 – Incorporação através de uma tabela de proteínas BC4r.xlsx - 3

```
token_id = dict()
id_token = dict()
BC4 = BC4.fillna('')
window_size = []

train_words = BC4.apply(lambda row: gen_tokens_of_line(
    token_id,
    id_token,
    window_size,
    row['Entry'],
    row['Protein names'],
    row['Subcellular location [CC]'],
    row['Interacts with'],
    row['Function [CC]'],
    row['Polymorphism'],
    row['Gene ontology (biological process)'],
    row['Gene ontology (cellular component)'],
    row['Gene ontology (molecular function)'],
    row['Domain [CC]']
), axis=1)

#Getting the highest window size
window_size = sorted(window_size, reverse=True)[0]
vocab_size = len(id_token)

embed_output = 100
cbow = Sequential()
cbow.add(Embedding(input_dim=vocab_size, output_dim=embed_output,
   input_length=window_size))
cbow.add(Lambda(lambda x: K.mean(x, axis=1), output_shape=(embed_output
   ,)))
cbow.add(Dense(vocab_size, activation='softmax'))
cbow.compile(loss='categorical_crossentropy', optimizer='rmsprop')
```

Código A.4 – Incorporação através de uma tabela de proteínas BC4r.xlsx - 4

```
def generate_context_word(word_row, token_id, window_size):
  dimension = len(token_id)
  word_vecs = []
  for word in word_row:
    word_vecs.append(token_id[word])

  for i in range(len(word_vecs)):
    inp = word_vecs[:i] + word_vecs[i+1:]
    inp = np.repeat(inp, window_size/len(inp))
    fill_size = window_size - len(inp)
    if (fill_size > 0):
      inp = np.hstack((inp, inp[:fill_size]))

    out = np.zeros(dimension)
    out[word_vecs[i]] = 1.0
    yield(
        inp.reshape(1, inp.shape[0]),
        out.reshape(1, dimension)
    )


for epoch in range(0, 8):
    loss = 0.
    print('\nStarting training for', len(BC4), 'rows')
    for i in range(0, len(BC4)):
      for x, y in generate_context_word(train_words[0], token_id,
          window_size):
        loss += cbow.train_on_batch(x, y)
      print('Finished row', i, 'of epoch', epoch)
    print('\tLoss:', loss, '\n')

weights = cbow.get_weights()[0]
print(weights.shape)

word_embeddings_df = pd.DataFrame(weights, index=[id_token[key] for key
   in sorted(id_token, reverse=True)])
word_embeddings_df.to_csv('word_embeddings.csv')
```

ANEXO B – DISTÂNCIAS DOS RESULTADOS NO PROT2VEC

Código B.1 – Utilização dos pontos resultados com a distância à palavra Angiogenic

```
import pandas as pd
import numpy as np

vectorscomp = dict()
vectorsprot = dict()
dotprot = dict()
BC2 = pd.read_excel('Data/BC2r.xlsx')
BC2WE = pd.read_csv('Data/BC2_word_embeddings.csv')


def one_hot_encoding_of_length(length):
    out = []
    values = list(dotprot.keys())
    labels = []
    for i in range(0, length):
        vec = np.zeros(length)
        vec[i] = 1.0
        out.append(vec)
        labels.append(values[i])
    return out, labels


def gen_vectors_complete(protein, vector):
    vectorscomp[protein] = np.array(vector)


def gen_vectors_prot(protein, vector):
    vectorsprot[protein] = np.array(vector)


proteins = BC2['Entry'].tolist()
vectorsdp = BC2WE[BC2WE['word'].isin(proteins)]
vectorsdp = vectorsdp.fillna(0.0)

BC2WE.fillna(0.0).apply(lambda row: gen_vectors_complete(row['word'],
   row[1:]), raw=True, axis=1)
vectorsdp.apply(lambda row: gen_vectors_prot(row['word'], row[1:]), raw=
   True, axis=1)

for prot, vec in vectorsprot.items():
    dotprot[prot] = vec.dot(vectorscomp['ANGIOGENIC'])

print(sorted(dotprot.items(), key=lambda x: x[1]))
```

ANEXO C – ÁRVORE DE DECISÃO SOBRE FUNÇÕES

Código C.1 – DCT relacionada às funções biológicas - 1

```
import pandas as pd
import numpy as np
import re
import graphviz
from sklearn import tree

def gen_tokens_functions(token_id, id_token, window_size, go_bio, name,
    interacts_with):
     words = go_bio.split('; ')
     interaction = re.sub(';', '', interacts_with)
     interaction = interaction.split(' ')
     for word in words:
         if token_id.get(word) is None:
             token_id[word] = len(token_id)
             id_token[len(token_id)] = word
     window_size.append(len(words))
     return [words, interaction, name]

def generate_train_words(bioprocesses, interactions, names, token_id,
    interest_functions):
     window = len(token_id)
     word_vecs = []
     has_angiogenesis = []
     angiogenesis_prots = []
     index = 0
     for word_row, name in zip(bioprocesses, names):
         row = np.zeros(window)
         row_has_angiogenesis = False
         for function in word_row:
             index = token_id[function]
             if index in interest_functions:
                 row_has_angiogenesis = True
                 angiogenesis_prots.append(name)
             else:
                 row[index] = 1.0
         word_vecs.append(row)
         has_angiogenesis.append(row_has_angiogenesis)
     for i, interacts in enumerate(interactions):
         for connection in interacts:
             if connection in angiogenesis_prots:
                 has_angiogenesis[i] = True

     return word_vecs, np.array(has_angiogenesis)
```

Código C.2 – DCT relacionada às funções biológicas - 2

```
BC2 = pd.read_excel('Data/BC2r.xlsx')
token_id = dict()
id_token = dict()
BC2 = BC2.fillna('')
window_size = []

tokens = BC2.apply(lambda row: gen_tokens_functions(
    token_id,
    id_token,
    window_size,
    row['Gene ontology (biological process)'],
    row['Entry'],
    row['Interacts with']
), axis=1)

bioprocesses = [item[0] for item in tokens]
interactions = [item[1] for item in tokens]
names = [item[2] for item in tokens]

window_size = sorted(window_size, reverse=True)[0]
vocab_size = len(id_token)

interest_functions = [token_id['angiogenesis [GO:0001525]'],
                      token_id['positive regulation of angiogenesis [GO
                          :0045766]'],
                      token_id['negative regulation of angiogenesis [GO
                          :0016525]'],
                      token_id['positive regulation of sprouting
                          angiogenesis [GO:1903672]'],
                      token_id['regulation of angiogenesis [GO:0045765]'
                          ],
                      token_id['positive regulation of cell migration
                          involved in sprouting angiogenesis [GO:0090050]
                          ']]

X, Y = generate_train_words(bioprocesses, interactions, names, token_id,
    interest_functions)

labels = sorted(token_id.items(), key=lambda x: x[1])
labels = [item[0] for item in labels]

angiogenesis_tree = tree.DecisionTreeClassifier(max_depth=5)
angiogenesis_tree.fit(X, Y)

dot_data = tree.export_graphviz(angiogenesis_tree, out_file=None,
   feature_names=labels, class_names=["Nao angiogenese", "Angiogenese"])
graph = graphviz.Source(dot_data)
graph.render("Angiogenesis")
```

ANEXO D – ÁRVORE DE DECISÃO SOBRE DADOS CLÍNICOS

Código D.1 – DCT relacionada aos dados clínicos

```
import pandas as pd
import numpy as np
import re
import graphviz
from sklearn import tree

Hist = pd.read_excel('Data/hist_data.ods')

Hist = Hist.dropna()

Hist['gender'] = Hist['gender'].map({'Male': 0, 'Female': 1})
X = Hist.drop('id', axis=1).drop('dead by melanoma', axis=1)
Y = Hist['dead by melanoma']

hist_tree = tree.DecisionTreeClassifier(max_depth=5)
hist_tree.fit(X, Y)

dot_data = tree.export_graphviz(hist_tree, out_file=None, feature_names=
    X.columns, class_names=["Vivo", "Morto"])
graph = graphviz.Source(dot_data)
graph.render("MortePorMelanoma")
```